UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Marcelo Buscioli Tenório

RECONHECIMENTO DE MODELOS DE PROBABILIDADE

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação.

Profa. Silvia Modesto Nassar, Dra.
(Orientadora)

Florianópolis, março de 2005.

RECONHECIMENTO DE MODELOS DE PROBABILIDADE

Marcelo Buscioli Tenório

Esta dissertação foi julgada adequada para obtenção do título de Mestre em Ciência da Computação. Área de Concentração Sistemas de Conhecimento e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Prof. Raul Sidnei Wazlawick, Dr. (Coordenador)

Banca Examinadora

_____

Profa. Silvia Modesto Nassar, Dra. (Orientadora)

_____

Prof. Paulo José de Freitas Filho, Dr.

_____

Prof. Mauro Roisenberg, Dr.

_____

Prof. Rogério Cid Bastos, Dr.

_____

Prof. Emil Kupek, Dr.

Dedico este trabalho a minha mãe
que pode acompanhar todo andamento,
e a meu pai que infelizmente só pode
vivenciar o início desta minha conquista.

Agradecimentos

Primeiramente agradeço a Deus por todas as providências e privilégios que ele me concedeu durante esta conquista. Sou muito agradecido pela minha vinda para a cidade de Florianópolis, pelos lugares que morei, pelos lugares que conheci, pela universidade que estudei e pelos amigos que fiz.

Agradeço a minha mãe Loide, ao meu pai Claudiom e aos meus irmãos, Marcio e Michele por tudo que fizeram por mim.

Agradeço aos professores com os quais tive contato durante o mestrado, especialmente a minha orientadora professora Silvia Nassar. Quero deixar neste parágrafo minha profunda admiração pela total atenção que a professora Silvia teve comigo no decorrer deste mestrado. Por tudo que fizemos juntos, desde as atividades dentro da universidade como disciplinas, projetos e a dissertação, até as atividades fora da universidade como as conversas em cafés e confraternizações em almoços.

Gostaria de agradecer aos professores Masanao Ohira, Rogério Cid Bastos e Paulo Freitas pelas contribuições nos meus momentos de dúvidas aos assuntos relacionados a dissertação. Quero também agradecer ao professor Masanao pelos empréstimos dos livros.

Também agradeço ao professor Paulo Freitas pelo convite a participar do projeto E&PRisk. Projeto desenvolvido pelo PerformanceLab - INE/UFSC em parceria ao Centro de Pesquisas da Petrobras (CENPES) na pessoa de Carlos Magno. Vale lembrar que o software implementado com base nesta pesquisa é um módulo do software E&PRisk.

Agradeço aos professores Mauro Roisenberg e Emil Kupek pelas contribuições oferecidas na avaliação desta dissertação.

Tenho alguns amigos em especial que não poderia esquecê-los. Agradeço a João Mafra, Nídia Bugra e família, por tudo que fizeram por mim desde quando cheguei em Florianópolis.

O meu muito obrigado a Renato Corrêa Vieira pela caminhada que fizemos por esta linda Florianópolis, pelos momentos de estudos, de descontração e até pelos desentendimentos. Apesar do Renato não ter acompanhado de perto a conclusão desta pesquisa, ele sempre esteve presente nas conversas por telefone.

Agradeço a Deus por ter conhecido André Luis Mello, ele veio de goiânia estudar Ciência da Computação na UFSC, mas não aguentou ficar longe de sua familia e voltou, hoje ele estuda na UFG. Bem, neste pouco tempo que ele ficou aqui, nossa amizade cresceu muito, fizemos viagens e ainda hoje, mesmo a distância, a gente se diverte muito. Valeu André.

O meu muito obrigado aos meus amigos Ygor Raphael, Guido Boin, Eduardo Feltrin, Cristiana Pasquini e Erika Feltrin. Amigos de Presidente Prudente, minha cidade natal, valeu pelos momentos em que estivemos nas praias e nos lugares maravilhosos de Floripa.

Agradeço a minha namorada Diane Almeida pela compreensão da minha ausência em alguns momentos, pelos conselhos, pelo amor e enfim, pela nossa caminhada juntos.

Agradeço a Cláudio Pereira Flores por todo tempo que dividimos, em disciplinas, em projetos e também nas atividades fora da universidade. Obrigado pelas conversas que tivemos e pela troca de experiência que até hoje temos, muito obrigado também pelos açais e cupuaçus que você, Cláudio, trouxe de Belém - PA.

Agradeço a Deus por ter conhecido meus atuais companheiros de apartamento, Fabio Pinheiro e Augusto Cesar. Valeu pelas boas músicas tocadas por Fábio em seu violão e pelos momentos de alegria nas sextas-feitas de sinuca. Apesar do Augusto sempre discordar das minhas opniões, sou grato por todas discussões que tivemos. Eu não poderia esquecer de agradecer as deliciosas comidas feitas em casa, em especial, o cuscuz nordestino. Enfim, valeu pelo tempo que passamos e que ainda estamos passando.

Valeu ter conhecido Carlos Tibiriça e Jaqueline Stumm, companheiros do mestrado, os mesmos também são orientados pela professora Silvia. Valeu por todas discussões sobre os mais diversos assuntos acadêmicos.

Enfim, este espaço para agradecimentos ficaria pequeno para eu descrever o quão agradecido estou a todos que direta ou indiretamente me ajudaram nesta caminhada.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas, Abreviações e Símbolos

IA	Inteligência Artificial
KS	Teste de aderência de Kolmogorov-Smirnov
AD	Teste de aderência de Anderson-Darling
MD	Módulo de Aderência
TN	Teste Tenório-Nassar
n	Tamanho da amostra de dados
na	Tamanho da amostra para o teste de aderência
K	Quantidade de classes
int	Resultado inteiro
max	Valor máximo
min	Valor mínimo
x_i	Valores da variável
$\bar{x}$	Média aritmética amostral
s^2	Variância amostral
s	Desvio padrão amostral
q_i	Quartil inferior
q_s	Quartil superior
m_d	Mediana (quartil do meio)
d_q	Desvio interquartílico
Ω	Espaço amostral
$p(x)$	Probabilidade de x
$F(x)$	Função de distribuição acumulada
$f(x)$	Função densidade de probabilidade
μ	Média populacional
e	Número Euler
π	Número *pi*

$\Re$	Conjunto dos números reais
σ^2	Variância populacional
χ^2	Distância qui-quadrado
H_0	Hipótese nula de um teste estatístico
H_1	Hipótese alternativa de um teste estatístico
valor-p	Probabilidade de significância
α	Nível de significância
z	Coeficiente z relativo a um determinado grau de confiança
me	Margem de erro
nc	Tamanho calculado
va	Valor de avanço entre as posições do vetor de dados
pp	Posição de parada no vetor, quando s_e_i atingir o valor cinco
pv	Posição dos valores
qv	Quantidade de valores
O_i	Freqüências observadas
E_i	Freqüências esperadas

Resumo

Reconhecimento de Padrões é uma área que objetiva descobrir e testar modelos matemáticos que possam identificar e explicar padrões em sinais de diversas fontes. Dados em grande quantidade, no formato texto, imagens, sons ou outras representações codificadas, são analisados para que sejam identificados padrões de relação.

Esta área tem estreita ligação com Redes Neurais Artificiais, Processamento de Imagens e com a Estatística. As técnicas de reconhecimento de padrão são utilizadas, entre outras aplicações, na detecção de ataques a redes de computadores, no reconhecimento de voz, imagem e impressões digitais e na interpretação de processos (fatos aleatórios). É este último caso o objeto de estudos da presente pesquisa.

Existem vários métodos estatísticos para reconhecimento de padrões referentes a modelos de distribuição de probabilidade, os quais são conhecidos como Testes de Aderência. Esta pesquisa tem o objetivo de propor soluções às dificuldades encontradas ao aplicar o Teste Qui-Quadrado de Aderência, tais como: o agrupamento dos dados em classes, valores discrepantes e a estimativa da distância qui-quadrado para cada classe.

A metodologia proposta utiliza os dados sem agrupá-los, faz crítica automática para identificação de valores discrepantes, calcula a distância qui-quadrado a partir das distribuições acumuladas e obtém a probabilidade de significância por aproximação da distribuição qui-quadrado pela normal.

A metodologia foi aplicada na implementação de um software na linguagem C++ para o sistema operacional Windows. Os resultados obtidos foram validados por comparações com os seguintes softwares: Statistica e Input Analyzer.

Palavras chave: Reconhecimento de Padrões, Distribuição de Probabilidade, Teste de Aderência.

Abstract

Pattern Recognition is an area that aims to discover and to test mathematical models which can identify and explain patterns in different sources. Large amounts of data, in text format, images, sounds or other codified representations are analyzed in order to identify related patterns.

This area has strong links with Artificial Neural Networks, Image Processing, and Statistics. Pattern recognition techniques are used, among other applications, in detection of computer network attacks, in voice, image, and fingerprint recognition, and process interpretation (random facts). The last case is the subject of this research.

There are many statistical methods of pattern recognition (probability distribution models), labeled as Goodness-of-fit Tests. The objective of this research is to propose solutions to the difficulties found in applying the Chi-square Goodness-of-fit Test, such as: data grouping in classes, outlying values and chi-square distance estimation for each class.

The proposed methodology uses data without grouping them, does automatic analysis for outlying value identification, calculates chi-square distance based on cumulative distributions, and obtains the significance probability by chi-square distribution approximation of the normal distribution.

The methodology was applied on software implementation in C++ on the Windows operating system. The results obtained were validated by comparison with these applications: Statistica and Input Analyzer.

Keywords: Pattern Recognition, Probability Distribution, Goodness-of-fit Test.

Capítulo 1

Introdução

No ano de 2005, afirmar que as informações contidas nos dados são vitais para uma organização qualquer, não é mais motivo de estudos e pesquisas para esta comprovação. Por outro lado, após a constatação da importância dos dados, pesquisas vêm sendo desenvolvidas no sentido de encontrar qual o melhor ou mais adequado método a ser aplicado e dessa forma recuperar o máximo possível das informações contidas nos dados. Em geral, os métodos de recuperação de informação são aplicados de acordo com o tipo de dado a ser processado e principalmente em relação ao objetivo a ser alcançado (BITTENCOURT, 2001).

É nesta vertente de estudos e pesquisas que áreas como engenharia e gestão do conhecimento, gestão das tecnologias da informação, inteligência computacional, entre outras, estão gradativamente descobrindo técnicas de reconhecer e aproveitar da melhor maneira possível os conhecimentos contidos em bases de dados.

As técnicas são as mais variadas, desde fundamentos matemáticos e estatísticos descobertos há muito tempo atrás, a novos estudos na área da Inteligência Artificial (IA), como por exemplo, na IA Simbólica: regras de produção, redes bayesiana e lógica difusa. Na IA Conexionista: redes neurais artificiais, na IA Evolucionária: programação evolucionária e algoritmos genéticos e ainda de forma Híbrida: fuzzy-bayesiana, estatística-IA e neuro-fuzzy (COHEN, 1995).

Com o advento dos computadores, as técnicas de descoberta e representação do conhecimento se tornaram muito úteis, visto que os computadores facilitaram a realização desses procedimentos que antes eram feitos manualmente ou por máquinas muito inferiores às atuais.

Focando ao assunto desta pesquisa, vale deixar claro que o processo de reconhecimento de modelos de distribuição de probabilidade foi realizado fazendo uso de métodos estatísticos. Neste contexto, a palavra reconhecimento tem o sentido de descobrir um ou mais modelos que melhor representam os dados analisados (BEZDEK, 1987).

Mas o que é um modelo? Segundo o dicionário Aurélio Eletrônico (1999), "modelo é aquilo que serve de base ou norma para a avaliação de qualidade. Modelo é um padrão ou um exemplo".

Especificamente nesta pesquisa, os modelos são probabilísticos, ou seja, existem funções matemáticas, as quais pode-se chamar de modelos, que representam probabilisticamente fatos aleatórios do dia a dia da humanidade.

Fazendo uma analogia entre os modelos probabilísticos e a definição da palavra modelo pelo dicionário Aurélio Eletrônico, pode-se dizer que um dos objetivos da presente pesquisa é descobrir padrões de dados, de maneira que possibilitem avaliar a qualidade dos dados, o quão eles se aproximam dos modelos probabilísticos.

Esta pesquisa tem a restrição de trabalhar com dados contínuos, mas acredita-se que a mesma também atenda a dados discretos. Dados contínuos estão presentes na vida das pessoas, por exemplo, tempo de espera em um ponto de ônibus, tempo total de ida e volta ao trabalho, duração de um curso de informática, tempo de perfuração de poços de petróleo, entre outros. Esses exemplos, chamados de fatos, podem ser representados por modelos probabilísticos. O problema então é descobrir entre vários modelos (funções de distribuição de probabilidade), qual melhor representa os dados coletados de determinado fato.

Se tal fato é representado por uma função de distribuição de probabilidade (modelo), o mesmo pode ser estudado com maior detalhes. É sabido que determinadas áreas de pesquisa, necessitam conhecer esse modelo, por exemplo, nas áreas de simulação, estatística e análise de decisão. Nessas áreas, as técnicas de análise são aplicadas de acordo com o padrão dos dados, por esse motivo, há necessidade de conhecer o modelo probabilístico que os dados seguem (NEWENDORP & SCHUYLER, 2000).

O reconhecimento de modelos de distribuição de probabilidade tem estreita ligação com o processo de reconhecimento de padrões. Sabe-se que o reconhecimento

de padrões tem sido utilizado até para analisar o comportamento do mercado de ações. Ao interpretar por determinado período as variáveis que mudam ou permanecem estáveis na bolsa de valores, está sendo possível diminuir os riscos das aplicações financeiras (COMPUTAÇÃO BRASIL, 2004).

É neste contexto de estudo que a presente pesquisa foi desenvolvida, a mesma oferece inovações à um dos métodos estatísticos tradicionais para reconhecimento de modelos probabilísticos.

Na estatística, os métodos para reconhecimento de modelos de probabilidade são conhecidos como teste de aderência. Existem vários testes de aderência, alguns com o comportamento melhor para dados discretos, outros para dados contínuos e outros para ambos os tipos de dados. Alguns exemplos de teste de aderência são: Teste Qui-quadrado, Teste de Kolmogorov-Smirnov, Teste de Lilliefors, Teste de Anderson-Darling, entre outros (JANKAUSKAS & MCLAFFERTY, 1995) (ROMEU, 2003, n.4, n.5 e n.6) (NIST/SEMATECH, 2005).

O objetivo de um teste de aderência é fazer o reconhecimento do padrão dos dados. Isto é feito de forma comparativa, escolhe-se um modelo e tendo-o como referência, verifica-se se os dados que estão sendo analisados seguem ou não o modelo escolhido.

Existem vários softwares ou funções estatísticas implementadas que realizam os testes de aderência, entre eles estão: Statistica, Input Analyzer, Simple Interactive Statistical Analysis, U-CARE, entre outros (STATISTICA, 2001) (INPUT, 2000) (SISA, 1997) (U-CARE, 2003).

1.1 Objetivo Geral

O objetivo desta pesquisa é desenvolver uma metodologia, alternativa às tradicionais, para reconhecimento de modelos de distribuição de probabilidade.

1.2 Objetivos Específicos

Para atingir o objetivo geral, alguns objetivos específicos foram delineados, são eles:

1) Investigar métodos quantitativos para o reconhecimento de modelos probabilísticos de dados amostrais;

2) Propor e implementar uma metodologia de reconhecimento de modelos probabilísticos;

3) Validar a metodologia proposta.

1.3 Justificativa

Seguindo a ordem de acontecimento dos fatos em relação ao uso dos dados, percebe-se uma certa evolução quando se trata da questão base de dados.

Observa-se que inicialmente houve a descoberta da utilidade dos dados, posteriormente, comprovado que os dados são úteis para extrair conhecimento, então como extrair esse conhecimento, e ainda mais, qual a melhor ou mais adequada forma para tal extração. Essa é uma breve seqüência do advento das linhas de pesquisa como Mineração de Dados, Descoberta de Conhecimento em Base de Dados e Reconhecimento de Padrões.

A presente pesquisa realiza um estudo que traz contribuições às linhas de pesquisa apontadas anteriormente. Fazem parte destas contribuições algumas inovações a um método estatístico tradicional para reconhecimento de modelos de distribuição de probabilidade. Buscou-se superar algumas limitações e dificuldades enfrentadas pelo método tradicional.

Após uma busca por trabalhos correlatos ao desta pesquisa, não foi encontrado nenhuma implementação que se assemelha à metodologia proposta e descrita a seguir.

Sabe-se, que em geral, os softwares existentes no mercado ou meio acadêmico realizam o teste qui-quadrado de aderência da forma que ele foi deduzido, ou seja, seguem os procedimentos tradicionais. A presente pesquisa descreve uma forma inovadora ao método tradicional para teste de aderência usando o teste qui-quadrado.

1.4 Estrutura da Dissertação

Esta dissertação tem a seguinte estrutura:

No capítulo 1 encontra-se a introdução, objetivo, objetivos específicos e a justificativa.

No capítulo 2 encontra-se a base conceitual, entre os conceitos da base estão, a organização de dados e variáveis, medidas descritivas, distribuições de freqüências, probabilidade, variáveis aleatórias e distribuições de probabilidades.

No capítulo 3 é descrito sobre testes de aderência, entre eles, Teste de Kolmogorov-Smirnov, Teste de Anderson-Darling e Teste Qui-quadrado de Aderência. Este capítulo descreve ainda o critério de decisão estatística pela probabilidade de significância, valor-p.

No capítulo 4 encontra-se a fundamentação teórica da metodologia proposta. No capítulo 5 encontra-se a matemática, resultados e validações da metodologia. O capítulo 6 apresenta a implementação do software. Para finalizar, o capítulo 7 exibe as considerações finais da presente pesquisa.

Capítulo 2

Organização dos Dados

Os dados são observações da realidade, sendo assim, os dados são essenciais para conhecer uma realidade. É processando dados que, estatísticas são realizadas, previsões são construídas, relatórios são montados e decisões são tomadas, ou seja, todos esses resultados finais iniciam-se pelos dados. Às vezes, quando não se têm dados, os mesmos são gerados segundo algum padrão especificado; em outros casos, procura-se iniciar o armazenamento de dados o mais rápido possível, para um posterior uso.

A exemplo de outras situações, os dados também precisam estar organizados para que haja um melhor aproveitamento dos mesmos. Neste sentido existem várias maneiras de se organizar e explorar esses dados, algumas delas são apresentadas a seguir.

2.1 Dados e Variáveis

Para exemplificar os conceitos a seguir, uma situação imaginária foi elaborada e é utilizada em todo desenvolvimento dos capítulos 2 e 3.

Com a intenção de verificar o perfil dos pós-graduados de um programa de pós-graduação, uma coleta de dados foi realizada com algumas variáveis tais como, número seqüencial do pós-graduado, sexo, nível de pós-graduação, idade com que concluiu a pós-graduação, duração do curso em anos e quantidade de disciplinas cursadas.

Antes de prosseguir com a explicação, o conceito dos termos população e amostra são necessários neste momento. População é o conjunto de elementos que formam o universo do estudo e que são passíveis de serem observados, sob as mesmas condições. Amostra é uma parte dos elementos de uma população (BARBETTA et al. 2004).

Retomando ao exemplo em questão, perfil dos pós-graduados, todos os alunos do programa de pós-graduação formam a população em estudo e os dados coletados para análise formam a amostra.

Após a aquisição dos dados, os mesmos precisam ser disponibilizados em algum formato. Costuma-se organizá-los na forma de tabela, esta tabela representa a amostra de dados a ser analisada. As colunas da tabela correspondem às variáveis levantadas e as linhas às realidades observadas.

A Tabela 2.1 mostra parte dos dados amostrais, perfil dos pós-graduados de um programa de pós-graduação, note que a mesma contém um total de cem realidades observadas.

Tabela 2. 1 - Parte da amostra de dados dos pós-graduados.

pós-graduado	**Sexo**	**nível**	**idade**	**duração**	**disciplinas**
1	Masculino	Doutorado	30	3,6	8
2	Feminino	Doutorado	31	4,2	8
3	Masculino	Doutorado	32	4,5	7
4	Feminino	Mestrado	26	2,0	8
5	Feminino	Mestrado	25	2,5	7
6	Masculino	Mestrado	26	3,1	7
7	Feminino	Doutorado	31	4,8	8
8	Masculino	Doutorado	29	3,5	7
9	Masculino	Doutorado	30	3,5	8
...	...	...	...	...	...
100	Masculino	Doutorado	30	3,4	8

Os dados podem ser observações de variáveis qualitativas ou de variáveis quantitativas. Para análise desses dados são aplicadas técnicas estatísticas diferenciadas de acordo com a classificação da variável.

Triola (1999, p.3) escreve que uma variável é dita qualitativa quando “seus dados podem ser separados em diferentes categorias que se distinguem por alguma característica não-numérica”. Completando a classificação das variáveis, o mesmo autor também define que uma variável é dita quantitativa quando “seus dados consistem em números que representam contagens ou medidas”.

Para ilustrar a classificação das variáveis e dos dados em termos do nível de mensuração, veja a Figura 2.1.

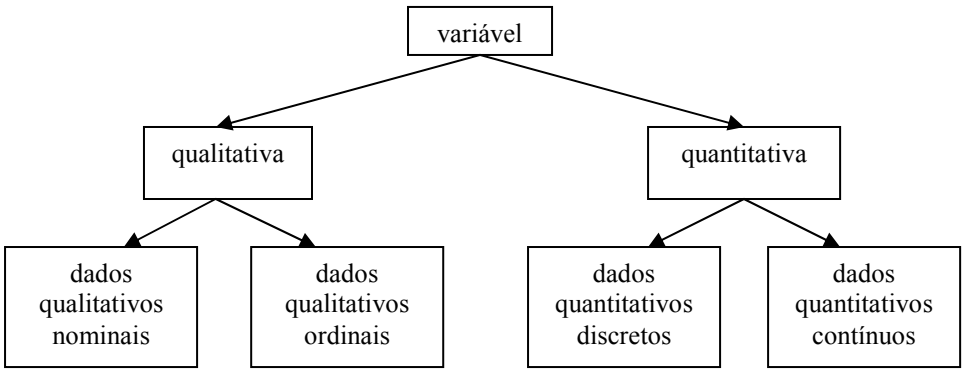


Figura 2. 1 - Classificação das variáveis e dos dados em termos do nível de mensuração.

Se os dados das variáveis qualitativas não podem ser colocados em ordem, esses são denominados nominais, se os dados podem seguir alguma ordem, são definidos como ordinais (ANDERSON & SCLOVE, 1986).

Se os dados das variáveis quantitativas resultam de um conjunto finito de valores possíveis, ou de um conjunto enumerável desses valores, esses são denominados discretos. Se os dados resultam de um número infinito de valores possíveis que podem ser associados a pontos em uma escala contínua de tal maneira que não haja lacunas ou interrupções, são definidos como contínuos (TRIOLA, 1999). É este último tipo de dados o objeto de estudos desta pesquisa.

Na Tabela 2.1 as variáveis, sexo e nível são exemplos de variáveis qualitativas, onde a variável, sexo, possui dados qualitativos nominais e a variável, nível, possui dados qualitativos ordinais. Por outro lado, as variáveis, idade, duração e disciplinas são quantitativas, sendo que a variável, disciplinas, possui dados discretos e as variáveis, idade e duração, possuem dados contínuos.

2.2 Distribuição de Freqüências

Continuando a exploração dos dados, o agrupamento desses dados é um dos primeiros passos a ser executado, este procedimento é chamado de distribuição de freqüências. Barbetta et al. (2004, p.53) escreve que "a distribuição de freqüências consiste na organização dos dados de acordo com as ocorrências dos diferentes resultados observados".

A elaboração de uma tabela de distribuição de freqüências varia de acordo com a classificação das variáveis e dos dados, observe a Tabela 2.2 que mostra uma distribuição de freqüências absolutas e relativas para variável qualitativa nível de pós-graduação.

Tabela 2. 2 - Distribuição de freqüências da variável nível de pós-graduação.

variável nível	**freqüência absoluta**	**freqüência relativa**
Doutorado	40	0,4
Mestrado	60	0,6
Total	**100**	**1,0**

Para variáveis qualitativas, tanto para dados nominais quanto para ordinais, a freqüência absoluta é obtida pela contagem dos elementos existentes em cada categoria, neste caso, Mestrado e Doutorado. A freqüência relativa é obtida pela divisão de cada valor da freqüência absoluta pela quantidade de realidades observadas, nesta situação, a quantidade de realidades observadas é igual a cem.

Para variáveis quantitativas, as quais se dividem em dados discretos e contínuos, a construção de uma tabela de distribuição de freqüências difere em parte das variáveis qualitativas.

Em posse de uma variável quantitativa com dados discretos, a elaboração da tabela de freqüências é análoga à construção da tabela para variáveis qualitativas. Por outro lado, tendo uma variável quantitativa com dados contínuos, a construção da tabela difere totalmente das variáveis qualitativas.

A montagem da tabela de freqüências para uma variável quantitativa com dados contínuos, tradicionalmente, é iniciada pela divisão da amplitude total dos dados em vários intervalos, denominados classes. A quantidade de classes a ser utilizada é uma escolha empírica, em geral, utiliza-se:

$$K = \sqrt{n} \qquad (2.1)$$

onde:

K = quantidade de classes,

n = tamanho da amostra de dados (quantidade de realidades observadas).

Cada intervalo de classe é obtido de forma arbitrária e subjetiva, a seguir segue uma sistemática que por vezes é utilizada:

$$int = (max - min) / K \tag{2.2}$$

onde:

int = resultado inteiro,

max = valor máximo observado na variável,

min = valor mínimo observado na variável,

o primeiro intervalo de classe é definido por:

$$[min;\ min + int) \tag{2.3}$$

o segundo intervalo por:

$$[min + int;\ min + (2*int)) \tag{2.4}$$

e assim sucessivamente até que o valor máximo (*max*) esteja contido no último intervalo de classe. Desta forma, todos os intervalos resultantes são de igual comprimento. Para concluir a construção da tabela é realizada a contagem dos elementos da amostra que se encontram no intervalo de cada classe.

Geralmente, este procedimento de construção da tabela de distribuição de freqüências de uma variável quantitativa com dados contínuos, gera uma tabela que não satisfaz as exigências para seu uso.

Concordando com Barbetta et al. (2004, p.60), essa insatisfação se dá pela forma subjetiva na construção da tabela, por exemplo, ao obter a quantidade de classes (*K*), não se leva em consideração a variabilidade dos dados e isto pode resultar uma tabela com muitas classes ou com poucas classes, o que não é bom, ou ainda classes com freqüência nula. Além da variabilidade dos dados, outros fatores como o tamanho da amostra e os objetivos da análise também devem ser levados em consideração na elaboração da tabela.

Por esse motivo, para esclarecer os demais conceitos, a Tabela 2.3 a seguir apresenta de forma simplificada (classes com amplitude unitária) a distribuição de freqüências para a variável duração, cujos dados são contínuos.

Tabela 2. 3 - Distribuição de freqüências da variável duração do curso em anos.

classes da variável duração	**freqüência absoluta**	**freqüência absoluta acumulada**
2,0 ⊢ 3,0	25	25
3,0 ⊢ 4,0	45	70
4,0 ⊢ 5,0	30	100

Note que diferentemente da Tabela 2.2, a Tabela 2.3 possui a coluna freqüência absoluta acumulada, onde o valor da primeira linha é o próprio valor da freqüência absoluta. A segunda linha é obtida pelo valor de sua freqüência absoluta somado ao valor da freqüência acumulada da linha anterior, assim sucessivamente até a última classe. O cálculo da freqüência acumulada também se aplica à freqüência relativa, resultando na freqüência relativa acumulada.

A distribuição de freqüências comumente é apresentada na forma de tabelas, em geral uma visualização gráfica torna-se mais sugestiva e atrativa em relação às tabelas, assim, a representação gráfica da distribuição é uma forma alternativa de apresentação dos dados.

Existem várias maneiras de representar graficamente uma tabela de freqüências, essas maneiras também variam de acordo com a classificação das variáveis. Para variáveis qualitativas existem, por exemplo, os gráficos de colunas, barras e setores. Para as variáveis quantitativas existem os gráficos diagrama de pontos, diagrama ramo-e-folhas e o histograma, este último é o de maior uso.

As Figuras 2.2 e 2.3 mostram os histogramas da Tabela 2.3, observe que os gráficos permitem a mesma análise dos dados que as tabelas, porém de forma mais clara.

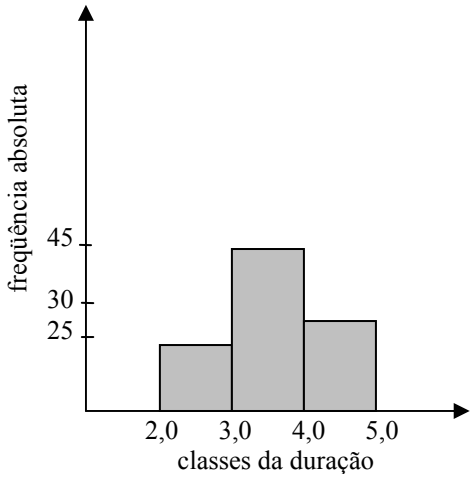


Figura 2. 2 - Distribuição de freqüências absolutas da variável duração do curso.

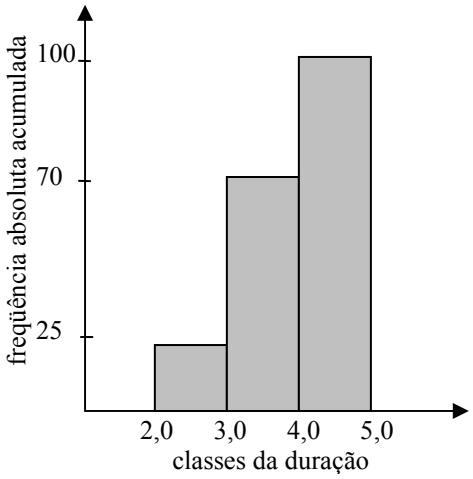


Figura 2. 3 - Distribuição de freqüências absolutas acumuladas da variável duração do curso.

Concluindo esta seção de distribuição de freqüências, torna-se relevante um complemento exploratório de dados a critério da classificação das variáveis. No âmbito da análise de dados, para variáveis qualitativas, as seguintes operações podem ser realizadas: elaboração da tabela de distribuição de freqüências e a criação dos gráficos. Sendo variáveis quantitativas, além da elaboração da tabela de freqüências e a criação dos gráficos, um outro recurso exploratório é possível, o uso das medidas descritivas, vale ressaltar que as medidas descritivas não se aplicam às variáveis qualitativas.

A seguir são apresentadas algumas das medidas descritivas, iniciando pela média aritmética e em seguida a mediana, moda, variância, desvio padrão e quartis.

2.3 Medidas Descritivas

O conceito de média aritmética, ou simplesmente média, é bastante familiar. Matematicamente, pode-se defini-la como um somatório dos valores dividido pela quantidade de valores somados. De modo geral, dado um conjunto de valores observados (n) de uma certa variável X, a média aritmética é dada por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (2.5)$$

onde:

x_i = valores da variável,

n = tamanho da amostra de dados.

Para descobrir qual a média de idade dos estudantes de pós-graduação, a média aritmética é a medida recomendada, ela resume o conjunto de dados em termos da posição central, ou de um valor típico. Porém não fornece informação suficiente sobre outros aspectos do conjunto, como assimetria e dispersão.

A mediana e a moda juntamente com a média são medidas que fornecem informações de assimetria. Quando um conjunto de valores estiver disposto em ordem crescente ou decrescente, a mediana é o valor central desse conjunto e a moda é o valor que ocorre com maior freqüência.

Um outro aspecto que por vezes se deseja observar em um conjunto de dados é em relação a sua dispersão. Para melhorar o resumo dos dados, pode-se apresentar junto à média, mediana e moda, uma medida da dispersão desses dados. Tanto a variância quanto o desvio padrão são medidas de dispersão em relação à média que em geral complementam o resumo dos dados.

A variância é definida como a média aritmética dos desvios quadráticos em relação à média, dada pela expressão:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad (2.6)$$

onde:

x_i = valores da variável,

$\bar{x}$ = média aritmética amostral,

n = tamanho da amostra de dados.

Como a variância de um conjunto de dados é calculada em função dos desvios quadráticos, sua unidade de medida equivale à unidade de medida dos dados ao

quadrado. Neste contexto, é mais comum trabalhar com a raiz quadrada positiva da variância, esta medida é conhecida como desvio padrão, o qual é expresso na mesma unidade de medida dos dados em análise. Então, o desvio padrão de um conjunto de valores é calculado por:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (2.7)$$

onde:

x_i = valores da variável,

$\bar{x}$ = média aritmética amostral,

n = tamanho da amostra de dados.

Ao comparar os desvios padrão de vários conjuntos de dados de uma mesma variável, pode-se avaliar quais se distribuem de forma mais (ou menos) dispersa. O desvio padrão será sempre positivo e ele extrai a informação de dispersão em relação à média dos valores observados, quanto mais dispersos estiverem os valores, maior o desvio padrão e vice-versa.

Os valores muito distantes da média dos dados são conhecidos como pontos discrepantes e eles podem descaracterizar algumas medidas descritivas dos dados tais como a média e a variância. Todavia, através das medidas quartis e extremos é possível identificar valores discrepantes. A seguir algumas definições para as medidas quartis e extremos são expostas.

Denomina-se de extremo inferior ao menor valor do conjunto de valores, isto é, $min(x_1, x_2, ..., x_n)$ e de extremo superior ao maior valor, isto é, $max(x_1, x_2, ..., x_n)$.

Denomina-se de primeiro quartil ou quartil inferior (q_i) o valor que delimita os 25% menores valores, de segundo quartil ou quartil do meio o valor que separa os dados em metades (50%) e de terceiro quartil ou quartil superior (q_s) o valor que separa os 25% maiores valores. O quartil do meio é a própria mediana, denotado por m_d. Os quartis inferior e superior também são conhecidos como quartis extremos.

A partir de uma amostra de dados de tamanho (n) e com os dados ordenados crescentemente, tem-se:

posição do $q_i : \frac{n+1}{4}$ posição do $m_d : \frac{n+1}{2}$ posição do $q_s : \frac{3(n+1)}{4}$

É importante ressaltar que os quartis (q_i, m_d e q_s) não são as posições (resultado das operações apresentadas acima) e sim os valores contidos nas respectivas posições, lembrando que os dados devem estar ordenados crescentemente.

A Figura 2.4 ilustra as divisões que os quartis e extremos realizam na distribuição dos dados.

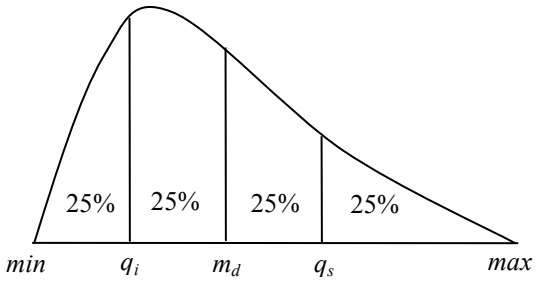


Figura 2. 4 - Divisão da distribuição dos dados em quartis e extremos.

Ressalta-se ainda que uma das funcionalidades dos quartis e extremos é detectar os pontos discrepantes. O desvio interquartílico ($d_q = q_s - q_i$) é muitas vezes usado como uma medida de dispersão e pode ser utilizado para verificar se há ou não pontos discrepantes nos dados em análise. São considerados pontos discrepantes os valores que ultrapassam $1{,}5d_q$ à direita do quartil superior e os valores que ultrapassam $1{,}5d_q$ a esquerda do quartil inferior.

2.4 Probabilidade

Para esclarecimento das questões probabilísticas, serão utilizados alguns experimentos aleatórios sobre os dados amostrais em questão, por exemplo, selecionar um pós-graduado e observar a ocorrência do sexo feminino.

Considerando a amostra de dados da Tabela 2.1, se um pós-graduado for selecionado ao acaso, não se pode afirmar a priori qual dos sexos, masculino (M) ou feminino (F), será selecionado. O resultado M ou F pode variar a cada seleção. Mas, ainda assim, há um certo padrão nos resultados, padrão esse que é evidenciado somente

após muitas repetições. Este fato notável é o fundamento da idéia de probabilidade (MOORE & MCCABE, 2002).

Em uma base de dados que possua metade do sexo feminino e a outra metade do sexo masculino, selecionar um pós-graduado e observar o sexo, apenas dois resultados são possíveis, masculino ou feminino. Para cada seleção, num total de 1.000, foi marcado a proporção da ocorrência sexo feminino. Na primeira seleção observou-se feminino, assim, a proporção de feminino começa em 1. Na segunda seleção observou-se masculino, reduzindo para 0,5 a proporção de feminino após duas seleções. Nas próximas três seleções observou-se um masculino seguido por dois femininos, portanto, a proporção de feminino após cinco seleções é de 3/5, ou 0,6. A proporção de seleção do sexo feminino é bastante variável no início, mas tende a se estabilizar a medida em que a quantidade de seleções aumenta. Aproximadamente, a partir da 150ª seleção, essa proporção tende para 0,5 e permanece aí. Assim, o valor 0,5 é a probabilidade de ocorrer sexo feminino, conforme mostra a Figura 2.5.

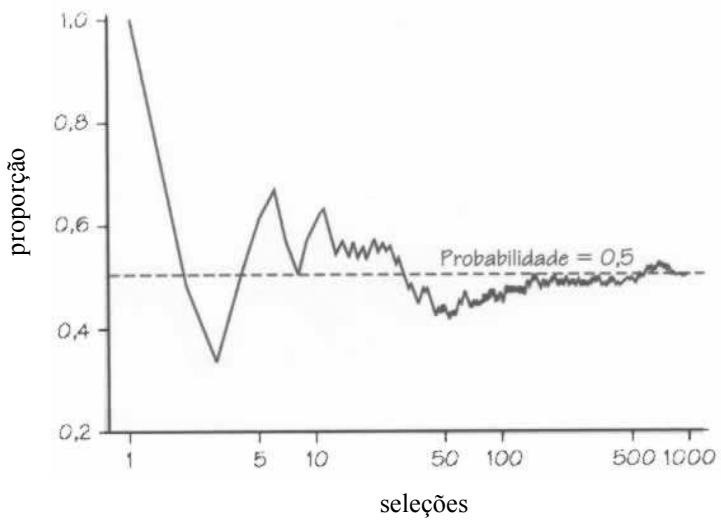


Figura 2. 5 - Proporção de seleção de um pós-graduado do sexo feminino.

Aleatório em estatística, não é sinônimo de casual ou acidental, e sim a descrição de um tipo de ordem que surge somente em longo prazo. Em nosso dia a dia, são encontrados freqüentemente fenômenos aleatórios, por exemplo, o tempo de espera em uma fila de banco. Raramente pode-se presenciar repetições suficientes do mesmo fenômeno aleatório, para observar a regularidade em longo prazo descrita pela

probabilidade. Essa regularidade pode ser observada na Figura 2.5. Em um prazo bastante longo, a proporção das seleções que dão sexo feminino é a probabilidade 0,5, isto significa que, ocorre metade das vezes em uma seqüência muito longa de provas. Essa é a idéia intuitiva de probabilidade (MOORE & MCCABE, 2002).

A concepção da probabilidade é empírica, isto é, baseia-se antes na observação do que na teoria. Para fixar uma probabilidade, deve-se observar o resultado de muitas provas para que, efetivamente, a probabilidade descreva o ocorrido nas provas.

Como exemplo de outro experimento aleatório, imagine uma moeda sendo lançada, você é capaz de dizer a proporção de caras?

Moore & McCabe (2002, p.166), apresentam alguns exemplos de pessoas que realmente têm, de fato, feito milhares de lançamentos de uma moeda e verificado a proporção de caras. Observe nos experimentos a seguir que a proporção de caras também permanece próxima ao 0,5:

> O naturalista francês Conde de Buffon (1707-1788) jogou uma moeda 4.040 vezes. Resultado: 2.048 caras, ou seja, uma proporção de 2048/4040 = 0,5069 caras.
> Por volta de 1900, o estatístico inglês Karl Pearson heroicamente jogou uma moeda 24.000 vezes. Resultado: 12.012 caras, ou seja, uma proporção de 0,5005.
> Quando estava prisioneiro dos alemães durante a Segunda Guerra Mundial, o matemático australiano John Kerrich jogou uma moeda 10.000 vezes. Resultado: 5.067 caras, ou seja, uma proporção de 0,5067.

A probabilidade estuda os processos que envolvem variabilidade e aleatoriedade. Para facilitar a análise, são construídos modelos matemáticos, normalmente, a partir de suposições sobre o processo, mas podem se basear também em dados observados no passado (BARBETTA et al., 2004).

Sobre probabilidade ou modelos probabilísticos, há dois aspectos a considerar. O primeiro é que intuitivamente as pessoas procuram tomar decisões em função dos fatos que têm maior probabilidade de ocorrer, por exemplo, se o céu está nublado, então há chance considerável de chover, recomenda-se levar um guarda-chuva ao sair de casa. O segundo aspecto é a incerteza inerente às decisões que podem ser tomadas sobre determinado problema, observe, por mais nublado que o céu esteja, pode não chover, ao

menos durante o período de tempo em que a pessoa estiver fora de casa. Se for possível quantificar a incerteza associada a cada fato, algumas decisões tornam-se mais fáceis.

"A teoria do cálculo de probabilidades permite obter uma quantificação da incerteza associada a um ou mais fatos e, portanto, é extremamente útil no auxílio à tomada de decisões" (BARBETTA et al., 2004).

Para concluir alguns conceitos sobre probabilidade, traga à mente o experimento de selecionar um pós-graduado e observar a ocorrência do sexo feminino. Neste caso, não se sabe exatamente qual será o sexo, apenas que será masculino (M) ou feminino (F), tem-se neste momento o espaço amostral desse experimento.

O conjunto de todos os possíveis resultados do experimento aleatório é chamado de espaço amostral e é denotado pela letra grega Ω. Neste caso:

$$\Omega = \{ M ; F \}.$$

Os elementos para se tomar alguma decisão podem corresponder a um conjunto de resultados (evento) associados ao experimento. É chamado de evento qualquer subconjunto do espaço amostral: A é um evento $\Leftrightarrow A \subseteq \Omega$ (BARBETTA et al., 2004). Observe um evento do experimento em questão:

$$A = \{ F \}.$$

Em geral, na estatística, quando se deseja entender algum experimento aleatório, o estudo é realizado com base em alguns eventos do experimento e suas chances de ocorrência. Este conceito é detalhado na próxima seção.

2.5 Variáveis Aleatórias

Para introduzir o conceito de variáveis aleatórias, atente-se para a seguinte colocação: antes de selecionar o pós-graduado, tente dizer qual será o sexo resultante. Não é possível afirmar a priori qual sexo ocorrerá, pois o resultado particular depende da chance de ocorrência e por isso é um fenômeno aleatório.

Segundo Moore & McCabe (2002, p.177), “uma variável aleatória é uma variável cujo valor é um resultado numérico de um fenômeno aleatório”.

Quando é selecionado um pós-graduado neste caso pode-se registrar F ou M. Em estatística, entretanto, quase sempre se está interessado em resultados numéricos tais como o número de sexo feminino em duas seleções de pós-graduados.

Seja X o número de ocorrência de sexo feminino após a seleção de dois pós-graduados. Se o resultado for o evento $\{ F ; F \}$, o valor de X será 2. Se o resultado for o evento $\{ M ; F \}$, o valor de X será 1. Os valores possíveis de X serão 0, 1 ou 2. Selecionando dois pós-graduados, X tomará um desses valores possíveis. É dito que X é uma variável aleatória, porque seus valores variam quando as seleções são repetidas.

Observe que há diferença entre o conjunto do espaço amostral Ω e o conjunto da variável aleatória X, porém existe uma relação (função) entre os dois conjuntos, conforme mostra a Figura 2.6 a seguir:

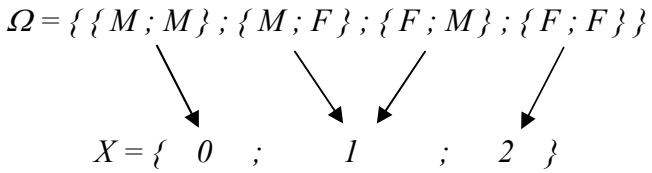


Figura 2. 6 - Relação entre os conjuntos espaço amostral e variável aleatória X.

Sendo assim, Barbetta et al. (2004, p.117), apresenta uma definição para variável aleatória, “formalmente, uma variável aleatória é uma função que associa elementos do espaço amostral ao conjunto de números reais”.

Vale destacar que existem casos em que os valores do espaço amostral coincidem com os valores da variável aleatória, por exemplo, o lançamento de um dado, neste experimento, os valores do espaço amostral formam o conjunto $\Omega = \{ 1 ; 2 ; 3 ; 4 ; 5 ; 6 \}$, igualando-se ao conjunto da variável aleatória X = número da face superior do dado, isto é $X = \{ 1 ; 2 ; 3 ; 4 ; 5 ; 6 \}$.

Conforme apresentadas anteriormente, as variáveis quantitativas podem possuir dados discretos ou contínuos. As variáveis aleatórias também são classificadas em discretas e contínuas.

De acordo com Triola (1999, p.93), “uma variável aleatória discreta ou admite um número finito de valores ou tem uma quantidade enumerável de valores” e “uma variável aleatória contínua pode assumir um número infinito de valores, e esses valores podem ser associados a mensurações em uma escala contínua, de tal forma que não haja lacunas ou interrupções”.

2.6 Distribuição de Probabilidades

Esta seção apresenta o conceito de distribuição de probabilidades para variáveis aleatórias discretas e contínuas.

A distribuição de probabilidades de uma variável aleatória discreta X é a descrição dos valores e das probabilidades associadas aos possíveis valores de X.

Ao selecionar dois pós-graduados, seja a variável aleatória discreta X = número de pós-graduados que cursaram oito disciplinas, com a ressalva de que a variável disciplinas assume apenas os valores sete e oito. Observe na Tabela 2.4 os eventos que compõem o espaço amostral desse experimento, os valores possíveis x da variável aleatória discreta X e suas probabilidades $p(x)$, respectivamente.

Tabela 2. 4 - Distribuição de probabilidades da variável aleatória discreta X = número de pós-graduados que cursaram oito disciplinas.

eventos	**valores possíveis** x	**probabilidades** $p(x)$
{ 7 ; 7 }	0	$1/4 = 0,25$
{ 7 ; 8 } , { 8 ; 7 }	1	$2/4 = 0,50$
{ 8 ; 8 }	2	$1/4 = 0,25$
Total		**1,00**

Ao selecionar dois pós-graduados e observar a ocorrência de terem cursado oito disciplinas, há quatro eventos (resultados) possíveis, todos com $1/4$ de probabilidade, ou seja, esses resultados são igualmente prováveis.

A variável X = número de pós-graduados que cursaram oito disciplinas tem os valores possíveis 0, 1 e 2. Mas esses valores não são igualmente prováveis. Conforme

apresentado pela Tabela 2.4, há apenas uma possibilidade de ocorrência $X = 0$, a saber, quando o resultado é *{ 7 ; 7 }*. Assim, $P(X = 0) = 1/4$, mas $X = 1$, pode ocorrer de duas maneiras diferentes *{ 7 ; 8 }* ou *{ 8 ; 7 }*, portanto, $P(X = 1) = 2/4$.

Segundo Barbetta et al. (2004, p.119), se X é uma variável aleatória discreta, com valores possíveis $x_1, x_2, ..., x_n$, então a distribuição de probabilidades de X pode ser dada pela função de probabilidade, que relaciona a cada valor possível x_i a sua probabilidade de ocorrência $p(x_i)$, ou seja:

$$p(x_i) = P(X = x_i), \text{ onde } i = 1, 2, ..., n \qquad (2.8)$$

As probabilidades $p(x_i)$ devem satisfazer duas exigências:

1) $p(x_i) \geq 0$,

2) $\sum_{i=1}^{n} p(x_i) = 1$.

Uma forma alternativa de representação da distribuição de probabilidades de uma variável aleatória é através da sua função de distribuição acumulada. Para as variáveis aleatórias discretas, a função de distribuição acumulada é definida por:

$$F(x) = P(X \leq x), \forall x \in \Re \qquad (2.9)$$

Assim, para todo $x \in \Re$, a função de distribuição acumulada descreve a probabilidade de ocorrer um valor menor ou igual a x.

Em geral, utiliza-se o gráfico em hastes para representar graficamente a distribuição de probabilidades de uma variável aleatória discreta, conforme apresentam as Figuras 2.7 e 2.8.

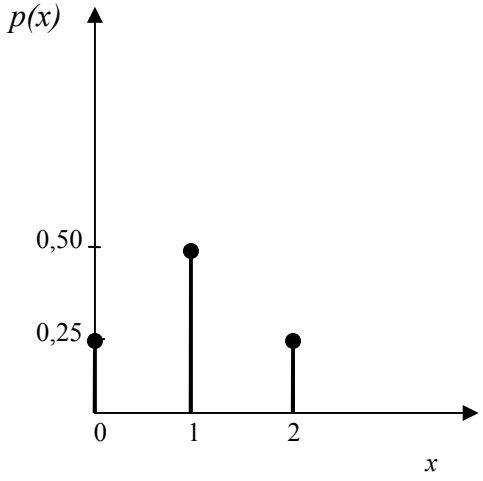


Figura 2. 7 - Representação gráfica da distribuição de probabilidades da variável aleatória discreta X = número de pós-graduados que cursaram oito disciplinas.

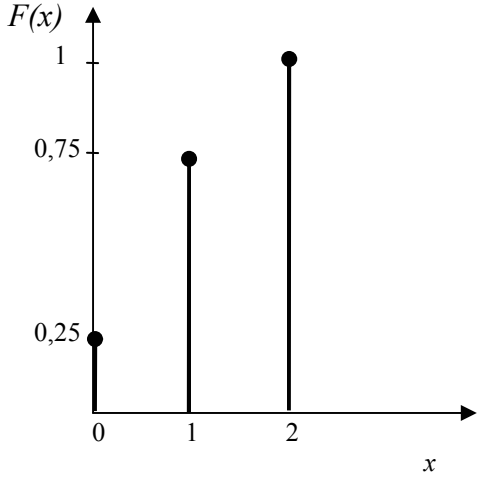


Figura 2. 8 - Representação gráfica da distribuição de probabilidades acumuladas da variável aleatória discreta X = número de pós-graduados que cursaram oito disciplinas.

No início desta seção foi definida que a distribuição de probabilidades de uma variável aleatória discreta X é a descrição dos valores e das probabilidades associadas aos possíveis valores de X. Entrando no âmbito das variáveis aleatórias contínuas, esta definição torna-se inviável, portanto, a distribuição de probabilidades de uma variável aleatória contínua X é descrita por uma função denominada função densidade de probabilidade. Veja pelos próximos parágrafos, o porque da diferença entre essas definições.

No exemplo anterior, ao selecionar dois pós-graduados, a variável aleatória X = número de pós-graduados que cursaram oito disciplinas é proveniente do seguinte espaço amostral: *{ { 7 ; 7 } ; { 7 ; 8 } ; { 8 ; 7 } ; { 8 ; 8 } }*. O modelo probabilístico atribui a probabilidade de $1/4$ a cada um dos quatro resultados possíveis, conforme apresentados pela Tabela 2.4.

Mudando o foco para a variável duração do curso, suponha que a variável X = conclusão do curso no intervalo de dois a três anos, admitindo como resultado qualquer número entre dois e três. O espaço amostral é agora todo o intervalo de números, ou seja:

$$\Omega = \{\forall x / 2{,}0 \leq x \leq 3{,}0\} .$$

Como é possível atribuir probabilidades a eventos tais como *{ 2,2 ≤ X ≤ 2,9 }*? Tal como no caso dos pós-graduados que cursaram oito disciplinas, é interessante, neste caso da conclusão do curso no intervalo de dois a três anos, se todos os resultados possíveis fossem igualmente prováveis. Mas não é possível atribuir probabilidade a cada valor individual de *X* e somá-los, porque há infinitos valores possíveis.

Para resolver este problema, é utilizada uma outra forma de atribuir probabilidades diretamente aos eventos, a probabilidade como área sob uma função de densidade. Um histograma é uma aproximação dessa função de densidade, conforme apresenta a Figura 2.9.

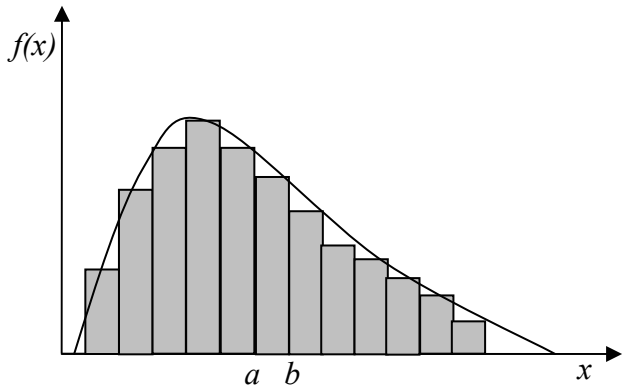


Figura 2. 9 - Um histograma como aproximação da função de densidade (distribuição de probabilidades) de uma variável aleatória contínua.

Segundo Montgomery & Runger (2003, p.74), se *X* é uma variável aleatória contínua, então a distribuição de probabilidades de *X* pode ser dada pela função densidade de probabilidade, tal que:

1) $f(x) \geq 0$,

2) $\int_{-\infty}^{\infty} f(x)dx = 1$,

3) $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx =$ área sob *f(x)* de *a* a *b* para qualquer *a* e *b*, como pode ser observado na Figura 2.9.

Para uma variável aleatória contínua, a função de distribuição acumulada é dada por:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx \text{ para } -\infty < x < \infty \qquad (2.10)$$

Como pode ser constatado, há uma certa similaridade entre as distribuições de probabilidades (funções) e as distribuições de freqüências (observadas) vistas em seções anteriores. Contudo, nas distribuições de probabilidades são mostrados os possíveis valores e não os valores efetivamente observados.

Aproveitando a colocação do parágrafo anterior, vale aqui destacar que para cada intervalo do histograma, a área da barra é igual à freqüência relativa (proporção) das medidas no intervalo. A freqüência relativa é uma estimativa da probabilidade da medida estar contida no intervalo. Similarmente, a área sob *f(x)* ao longo de qualquer intervalo é igual à probabilidade verdadeira da medida estar contida no intervalo. Isto pode ser evidenciado pela Figura 2.9, na qual as barras constituem o histograma e a curva que acompanha o histograma representa a densidade.

No exemplo da variável aleatória discreta *X* = número de pós-graduados que cursaram oito disciplinas, a distribuição de probabilidades foi construída empregando um conhecimento empírico para o cálculo das probabilidades envolvidas. A teoria da probabilidade oferece alguns modelos probabilísticos teóricos que auxiliam na obtenção da distribuição de probabilidades.

No caso das variáveis aleatórias discretas, são exemplos de modelos teóricos: distribuição de Bernoulli, binomial, hipergeométrica e Poisson. Para variáveis contínuas, a Tabela 2.5 a seguir, apresenta alguns modelos teóricos de distribuição de probabilidades.

Tabela 2. 5 - Funções de distribuição de probabilidades para variáveis contínuas.

(*f(x)* = função densidade de probabilidade e *F(x)* = função de distribuição acumulada).

Uniforme	
$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{se } a \le x \le b, \\ 0 & \text{caso contrário.} \end{cases}$	$F(x) = \begin{cases} 0 & \text{se } x < a, \\ \dfrac{x-a}{b-a} & \text{se } a \le x \le b, \\ 1 & \text{se } b < x. \end{cases}$

onde:

$a = min, \forall a \in \Re$,

$b = max, \forall b \in \Re, b > a$.

Exponencial	
$f(x) = \begin{cases} \dfrac{1}{\mu} e^{-x/\mu} & \text{se } x \ge 0, \\ 0 & \text{caso contrário.} \end{cases}$	$F(x) = \begin{cases} 1 - e^{-x/\mu} & \text{se } x \ge 0, \\ 0 & \text{caso contrário.} \end{cases}$

onde:

μ = média populacional, $\mu > 0$,

e = número Euler.

Normal	
$f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$, $\forall x \in \Re$.	Sem função *F(x)* definida.

onde:

π = número *pi*,

$\Re$ = conjunto dos números reais,

σ^2 = variância populacional,

e = número Euler,

μ = média populacional, $\mu \in (-\infty , \infty)$.

Lognormal

$$f(x)=\begin{cases}\dfrac{1}{x\sqrt{2\pi\sigma^2}}e^{-(\ln x-\mu)^2/2\sigma^2} & \text{se } x>0,\\ 0 & \text{caso contrário.}\end{cases}$$

Sem função *F(x)* definida.

onde:

π = número *pi*,

σ^2 = variância populacional,

e = número Euler,

μ = média populacional, $\mu \in (-\infty, \infty)$,

ln x = logaritmo natural de *x*.

Triangular

$$f(x)=\begin{cases}\dfrac{2(x-a)}{(b-a)(c-a)} & \text{se } a \le x \le c,\\ \dfrac{2(b-x)}{(b-a)(b-c)} & \text{se } c \le x \le b,\\ 0 & \text{caso contrário.}\end{cases}$$

$$F(x)=\begin{cases}0 & \text{se } x<a,\\ \dfrac{(x-a)^2}{(b-a)(c-a)} & \text{se } a \le x \le c,\\ 1-\dfrac{(b-x)^2}{(b-a)(b-c)} & \text{se } c<x \le b,\\ 1 & \text{se } x>b.\end{cases}$$

onde:

$a = min, \forall a \in \Re$,

$c = moda, \forall c \in \Re$,

$b = max, \forall b \in \Re, a < c < b$.

As funções dos modelos teóricos de distribuição de probabilidades fazem uso dos chamados parâmetros populacionais (θ), como a média populacional (μ), variância populacional (σ^2), desvio padrão populacional (σ) e outros. O ideal é que esses parâmetros sejam obtidos pelos dados populacionais, mas em geral isso não é possível. Entretanto, os parâmetros podem ser estimados a partir de uma amostra de dados da população em estudo, esta estimativa é denotada por $\hat{\theta}$.

Os parâmetros populacionais são estimados por estimadores tais como, média amostral ($\bar{x}$), mediana, variância amostral (s^2), desvio padrão amostral (s) e outros (CAZORLA, 2004). Fazendo o uso de uma amostra para estimar parâmetros populacionais, surgem alguns questionamentos, como:

Quais as características desejáveis para um bom estimador?

Como decidir que um estimador é melhor que outro?

Isto significa que é necessário estipular alguns critérios que respondam a estes questionamentos. A seguir são apresentados os conceitos que os estimadores precisam atender para resolução dos problemas de estimação.

Meyer (1984), Bowker & Lieberman (1972) e Cazorla (2004) escrevem as seguintes definições:

1) Um estimador T é dito não-tendencioso de θ, se $E(T) = \theta$.
2) Seja T um estimador não-tendencioso de θ. É dito que T é um estimador não-tendencioso e de variância mínima de θ, se para todos os estimadores T^* tais que $E(T^*) = \theta$, tiver $V(T) \leq V(T^*)$. Para este caso, outros autores chamam de estimador eficiente.
3) Um estimador T é consistente se $lim_{n \to \infty} P(|T - \theta| > \varepsilon) = 0$ para todo $\varepsilon > 0$. Para este caso, outros autores chamam de estimador coerente.
4) Um estimador é chamado de suficiente se contêm o máximo possível de informação com referência ao parâmetro por ele estimado.

Concluindo este capítulo, deseja-se a partir de um conjunto de valores observados de uma variável aleatória, descobrir qual modelo teórico de probabilidade, e seus parâmetros, pode representar mais adequadamente esse conjunto de valores. Este processo de reconhecimento de modelos de distribuição de probabilidade é um teste estatístico chamado Teste de Aderência. O próximo capítulo traz explicações sobre alguns testes de aderência e o uso da probabilidade de significância (valor-p) na tomada de decisão estatística.

Capítulo 3

Teste de Aderência

Os testes estatísticos são divididos em dois tipos, os testes paramétricos e os testes não paramétricos. Nos testes paramétricos é suposto que os dados seguem determinada distribuição de probabilidades, em geral, a distribuição normal. Porém, os testes não paramétricos são utilizados quando as suposições para se aplicar os testes paramétricos não são satisfeitas.

Montgomery & Runger (2003, p.347) colocam que geralmente, procedimentos não paramétricos não utilizam toda a informação fornecida pela amostra de dados. Como resultado, para uma mesma amostra, um procedimento não paramétrico será menos eficiente que o procedimento paramétrico correspondente quando a população em questão for normal. Essa perda de eficiência é refletida por uma necessidade de um tamanho maior de amostra para o procedimento não paramétrico do que o requerido pelo procedimento paramétrico. Por outro lado, a perda de eficiência não é geralmente grande e freqüentemente a diferença no tamanho da amostra é muito pequena.

Os testes de aderência fazem parte dos testes não paramétricos. O objetivo dos testes de aderência é verificar se os dados de uma amostra aderem (seguem) a uma determinada distribuição teórica, daí o nome teste de aderência. Essa distribuição teórica pode ser uma distribuição de probabilidades clássica (normal, exponencial, entre outras) ou proporções definidas especificamente para o problema (Lei de Mendel para ervilhas lisas e rugosas). São exemplos de testes de aderência, Teste Qui-quadrado, Teste de Kolmogorov-Smirnov, Teste de Anderson-Darling e Teste de Lilliefors (ROMEU, 2003, n.4, n.5 e n.6) (BARBETTA et al., 2004).

3.1 Teste de Aderência de Kolmogorov-Smirnov

Este teste é bastante utilizado em ocasiões que se deseja verificar a aderência de um conjunto de valores em relação a uma distribuição de probabilidades especificada. Embora seja possível aplicar outros testes, como o teste qui-quadrado de aderência descrito na próxima seção, nestas condições, geralmente é melhor aplicar o chamado teste de aderência de Kolmogorov-Smirnov (KS) (BARBETTA et al., 2004).

O teste KS diferentemente de outros testes de aderência, utiliza em seu método, a função de distribuição na sua forma acumulada. Seja *F(x)* a função de distribuição acumulada, com parâmetros especificados, para a qual se deseja verificar a aderência dos dados. As hipóteses são:

H_0: os dados provêm de *F(x)* (há aderência),
H_1: os dados não provêm de *F(x)* (não há aderência).

Sejam as distribuições acumuladas: a empírica *S(x)* e a teórica *F(x)*. Para cada elemento da amostra, obtém-se a diferença absoluta entre essas duas distribuições. A estatística do teste é a diferença absoluta máxima, *D* (ROMEU, 2003, n.6). O método é descrito a seguir:

1) Define-se *S(x)* para cada valor x_i ($i = 1, 2, ..., n$) como:

$$S(x_i) = \frac{qv \leq x_i}{n} \qquad (3.1)$$

onde:

qv = quantidade de valores,
n = tamanho da amostra de dados,
x_i = um valor da amostra.

2) Obtém-se para cada valor x_i ($i = 1, 2, ..., n$), o valor teórico *F(x)*, calculado pela função de distribuição acumulada, especificada em H_0.
3) Verifica-se a discrepância entre *S(x)* e *F(x)* através das diferenças absolutas entre $F(x_i)$ e $S(x_i)$, e entre $F(x_i)$ e $S(x_{i-1})$, para $i = 1, 2, ..., n$.
4) Calcula-se a estatística do teste, *D*, em termos da amostra em análise:

$$d = \max\{|F(x_i) - S(x_i)|, |F(x_i) - S(x_{i-1})|\} \qquad (3.2)$$

5) Uma vez identificada a distância máxima *d*, elabora-se as hipóteses de decisão. O processo de decisão está descrito pela seção 3.3 a seguir.

Segundo Nist/Sematech (2005), uma das características atrativas deste teste é que a distribuição da estatística do teste não depende de uma função de distribuição base, como acontece com o teste qui-quadrado. Uma outra vantagem é que ele é um teste exato, ou melhor, não depende de um tamanho adequado da amostra para a aproximação ser válida, como também acontece com o teste qui-quadrado.

A despeito das vantagens, o teste KS tem algumas limitações que devem ser consideradas (NIST/SEMATECH, 2005):

1) Aplica-se somente em distribuições contínuas;
2) Tende a ser mais sensível na faixa central da distribuição do que nos extremos;
3) Talvez a mais grave limitação é que a distribuição precisa ser completamente especificada. Isto é, a estimação dos parâmetros utilizando os dados amostrais, nem sempre levam a resultados satisfatórios.

Devido às limitações 2 e 3, muitos pesquisadores preferem usar o teste de aderência de Anderson-Darling (AD). Entretanto, o teste AD é aplicável somente em algumas distribuições específicas.

O teste AD é uma variação do teste KS. É dado mais importância para os extremos da distribuição do que no teste KS. O teste KS tem a distribuição livre no sentido que os valores críticos não dependem da distribuição específica sendo testada. O teste AD faz uso da distribuição específica testada no cálculo dos valores críticos. Isto tem a vantagem de permitir um teste mais sensível e a desvantagem que os valores críticos devem ser calculados para cada distribuição.

Um outro teste é o teste de aderência de Lilliefors, este teste também é uma variação do teste KS. Segundo Barbetta et al. (2004), ele é usado para verificar a aderência dos dados a uma distribuição normal qualquer, isto é, seus parâmetros (média e desvio padrão) são calculados com base na amostra.

O teste de Lilliefors é bastante utilizado para avaliar se é possível aplicar um teste paramétrico que supõe a distribuição normal.

3.2 Teste Qui-quadrado de Aderência

O teste qui-quadrado de aderência pode ser aplicado quando se está estudando dados distribuídos em classes, essas classes podem provir de uma variável qualitativa ou mesmo de uma variável quantitativa com dados discretos ou contínuos. No caso de uma variável quantitativa, especificamente com dados contínuos, tradicionalmente, costuma-se classificar os dados, formando uma distribuição de freqüências com dados agrupados, conforme é colocado no capítulo 2 (ROMEU, 2003, n.4).

A utilização do teste qui-quadrado de aderência se dá quando há interesse em verificar se as freqüências observadas nas K diferentes classes $(O_i, i = 1,2,...,K)$ são significativamente distintas de um conjunto de K freqüências esperadas $(E_i, i = 1,2,...,K)$. As hipóteses são:

$H_0 : O_i = E_i$ para todo $i = 1,2,...,K$,

$H_1 : O_i \neq E_i$ para algum $i = 1,2,...,K$.

A estatística desse teste, chamada de distância χ^2 é uma espécie de medida de distância entre as freqüências observadas (amostra) e as freqüências esperadas (distribuição de probabilidades teórica) de cada classe. Sua expressão é dada por:

$$\chi^2 = \sum_{i=1}^{K} \left[\frac{(O_i - E_i)^2}{E_i} \right]. \qquad (3.3)$$

onde:

O_i = freqüências observadas,

E_i = freqüências esperadas,

K = quantidade de classes.

Havendo aderência (H_0 verdadeira), as freqüências observadas devem ficar próximas das esperadas, acarretando um valor pequeno para χ^2: as variações encontradas seriam apenas casuais. Contudo, se não houver aderência (H_1 verdadeira),

diferenças entre freqüências observadas e esperadas poderão ser grandes, resultando em um valor grande para χ^2: sendo pouco provável que as variações tenham sido casuais.

A distancia χ^2 segue aproximadamente uma distribuição qui-quadrado com *K-1* graus de liberdade, onde *K* = número de classes. Esta definição aponta a distribuição qui-quadrado como a distribuição de referência do teste.

Uma das características atrativas deste teste é que ele pode ser aplicado a qualquer distribuição de uma variável (discreta ou contínua) para a qual possa ser obtida a função de distribuição acumulada.

Segundo Nist/Sematech (2005), o teste de aderência qui-quadrado é aplicado em dados agrupados (dados dispostos em classes). Isto não é de fato uma restrição desde que para dados não agrupados você possa obter um histograma ou uma tabela de freqüência antes de gerar o teste qui-quadrado. Entretanto o valor da estatística do teste qui-quadrado é dependente de como os dados estão agrupados. Outra desvantagem do teste qui-quadrado é que ele requer um tamanho de amostra suficiente para que a aproximação qui-quadrado seja válida.

Para um melhor entendimento desse teste, tente responder a seguinte pergunta: A idade dos pós-graduados segue uma distribuição exponencial?

Uma forma visual de resposta a esta pergunta é construindo o histograma da distribuição de freqüências da variável idade dos pós-graduados e verificando graficamente se segue ou não a uma distribuição exponencial. Mas essa forma visual pode não ser uma alternativa muito confiável, porém, há a necessidade de um teste matemático. São em situações deste tipo que os testes de aderência são aplicados.

O raciocínio do teste qui-quadrado de aderência, para a pergunta em pauta, se inicia com a construção da tabela de distribuição de freqüências (O_i) e com a elaboração da distribuição de probabilidades do modelo exponencial (E_i) em cada classe, para então submeter à expressão 3.3 e obter o valor da distância χ^2.

É importante ressaltar que não são todos os valores de cada classe que participam do cálculo da distância χ^2, somente o valor médio da classe é utilizado no cálculo.

Lembrando que quanto menor o valor χ^2, melhor a aderência dos dados observados em relação à distribuição exponencial, ou seja, é provável que a idade dos pós-graduados siga a uma distribuição exponencial.

3.3 Valor-p

Em qualquer teste de aderência, para decidir estatisticamente se há aderência dos dados a uma distribuição específica, é necessário elaborar as hipóteses de decisão e para isto existem duas abordagens, a abordagem clássica que usa tabelas estatísticas e uma outra abordagem que usa o valor-p.

A abordagem clássica não será apresentada pelo fato da mesma não ser utilizada nesta pesquisa, porém, a abordagem do valor-p está detalhada a seguir.

Dada uma hipótese H_0 e um conjunto de dados amostrais, o valor-p reflete a plausibilidade de se obter tais resultados no caso de a hipótese H_0 ser, de fato, verdadeira. O valor-p é chamado de probabilidade de significância e é obtido a partir dos dados amostrais. Um valor-p muito pequeno constitui evidência contra a hipótese H_0 e a favor da hipótese H_1 (TRIOLA, 1999).

Na realização de uma pesquisa quantitativa que utiliza a estatística para analisar os dados, quando se deseja confirmar ou refutar alguma hipótese, é comum estabelecer, ainda na fase de planejamento do estudo, a probabilidade tolerável de incorrer no erro de rejeitar H_0, quando H_0 é verdadeira. Este valor é conhecido como nível de significância do teste e é designado pela letra grega α. É comum adotar nível de significância de 5%, isto é, $\alpha = 0,05$. Mas quando se deseja maior segurança ao afirmar H_1, pode-se adotar níveis de significância menores, como $\alpha = 0,01$.

Retomando o exemplo que questiona se a idade dos pós-graduados segue a uma distribuição exponencial, têm-se as seguintes hipóteses a serem testadas:

H_0: a idade segue a uma distribuição exponencial,
H_1: a idade não segue a uma distribuição exponencial.

Pelo teste de aderência qui-quadrado, depois de ter calculado a distância χ^2, como então obter o valor-p? O valor-p é obtido pela expressão a seguir:

$$valor\text{-}p = 1 - P. \qquad (3.4)$$

onde:

P = probabilidade até a distância χ^2.

Para dados discretos, *P* é o somatório das probabilidades de cada valor *x*. Para dados contínuos, *P* é a área sob a curva. Em ambos os casos a probabilidade *P* é calculada até a distancia χ^2. Observe a Figura 3.1 a seguir, na qual há uma ilustração da expressão 3.4 para o cálculo do valor-p.

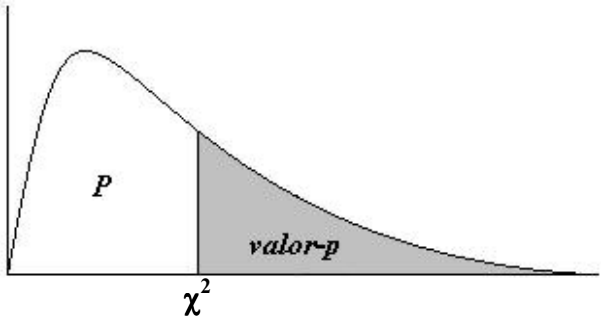


Figura 3. 1 - Ilustração do valor-p usando o modelo qui-quadrado com *K-1* graus de liberdade.

Para concluir o raciocínio, após estabelecer o nível de significância α, tem-se a seguinte regra geral de decisão do teste estatístico:

valor-p > α ➔ aceita H_0,
valor-p ≤ α ➔ rejeita H_0.

Finalizando o exemplo em questão, se valor-p > α então a idade dos pós-graduados segue a uma distribuição exponencial (aceita H_0), caso contrário, não segue tal distribuição (rejeita H_0).

Capítulo 4

Metodologia Proposta

Este capítulo descreve as inspirações conceituais que fundamentam a metodologia proposta, tendo o objetivo desta pesquisa como alvo a ser alcançado. Para isto, foi necessário tratar alguns pontos negativos que o teste qui-quadrado de aderência apresenta.

Esta metodologia proposta é uma forma alternativa, ao teste qui-quadrado de aderência tradicional, para reconhecimento de modelos de distribuição de probabilidade.

A Figura 4.1 a seguir mostra um diagrama dos métodos citados anteriormente para teste de aderência. Observe que existem duas especializações do teste KS, são elas: teste AD e teste Lilliefors. A presente pesquisa descreve uma especificação do teste qui-quadrado, chamada de teste Tenório-Nassar (TN).

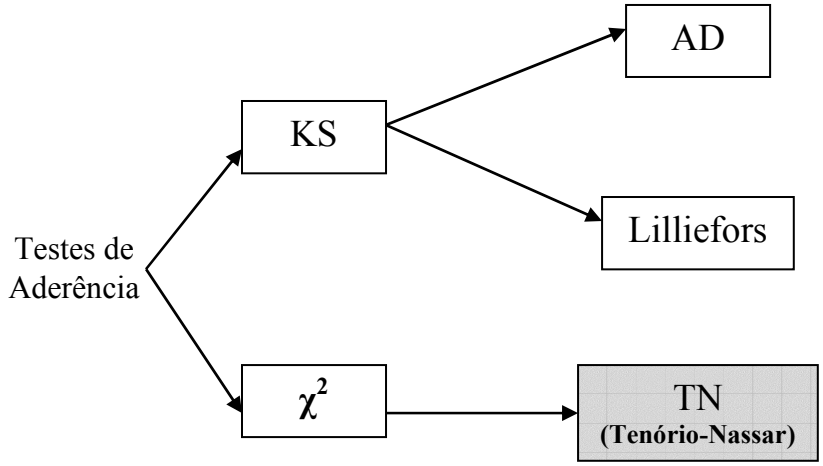


Figura 4. 1 - Especialização do teste qui-quadrado.

4.1 Fundamentos da Metodologia Proposta

No primeiro momento do processo de reconhecimento de modelos probabilísticos, é necessário escolher um teste estatístico. Em geral, os testes de aderência, como o teste de Kolmogorov-Smirnov e o teste de Lilliefors verificam a aderência de um conjunto de valores em relação a uma distribuição de probabilidades especificada.

Uma distribuição de probabilidades especificada é uma distribuição com seus parâmetros conhecidos. Nessas condições, os testes Kolmogorov-Smirnov e Lilliefors são alternativas mais poderosas (NIST/SEMATECH, 2005).

Uma distribuição com parâmetros desconhecidos, os mesmos precisam ser estimados (calculados) a partir da amostra de dados em questão. É neste segundo caso que a presente pesquisa se enquadra.

Tendo em vista que os dados a serem analisados podem provir de uma amostra qualquer, ou seja, não havendo certeza de que os dados sigam uma específica distribuição de probabilidades (parâmetros desconhecidos), o teste qui-quadrado de aderência é a alternativa recomendada, esse é um dos motivos pela escolha do teste qui-quadrado de aderência.

O teste qui-quadrado pode ser aplicado a distribuições discretas e contínuas. Os outros testes citados anteriormente (KS, AD e Lilliefors) são aplicados somente a distribuições contínuas. Apesar desta pesquisa se dedicar somente às distribuições contínuas, vale destacar que ela pode ser estendida para distribuições discretas. Este é outro motivo pela escolha do teste qui-quadrado.

Antes de iniciar o procedimento do teste qui-quadrado de aderência, recomenda-se avaliar a dispersão dos dados em relação à média. A presente metodologia sugere que esta avaliação de dispersão seja realizada através das medidas descritivas quartis e extremos. Conforme está escrito no capítulo 2, com estas medidas é possível detectar pontos que estão muito distantes da média (discrepantes) e que prejudicam o processo de reconhecimento dos modelos. Para melhorar a qualidade da informação dos dados, estes pontos discrepantes devem ser eliminados.

Para realizar o processo de reconhecimento de modelos usando o teste tradicional qui-quadrado, visto que os dados são contínuos, é necessário construir uma tabela de distribuição de freqüências, ou seja, agrupar os dados.

Nesta etapa da exploração dos dados, já se encontra a primeira desvantagem do método, a dificuldade é especificar o número de classes a ser utilizado no agrupamento dos dados (tabela de distribuição de freqüências).

Uma tabela de freqüências com poucas classes apresenta a distribuição de forma bastante resumida, podendo deixar de evidenciar algumas características relevantes. Por outro lado, uma tabela com muitas classes pode não realçar aspectos relevantes da distribuição de freqüências (BARBETTA et al., 2004).

Para resolver esta dificuldade, optou-se por analisar os dados brutos, ou seja, sem agrupá-los em classes. Sendo assim, para gerar as distribuições (observada e esperada), o uso da freqüência acumulada (observada) e da função de distribuição acumulada (esperada) tornou-se obrigatório, pois a forma acumulada permite trabalhar com os valores individualmente.

Em alguns momentos no decorrer do teste, este acúmulo é apresentado na forma relativa, por exemplo, na construção dos gráficos histograma e curva da distribuição teórica. Em outros momentos, por exemplo, no cálculo da distância χ^2, o acúmulo é utilizado na sua forma absoluta.

Ainda sobre o assunto da metodologia não trabalhar com dados agrupados em classes, uma outra opção para gerar as distribuições (observada e esperada), seria o uso da freqüência observada (não acumulada) juntamente com a função densidade de probabilidade (esperada), mas nesta opção os dados obrigatoriamente precisam estar agrupados, levando novamente ao problema da especificação do número de classes a ser utilizado no agrupamento dos dados. A obrigatoriedade do agrupamento dos dados pode ser melhor compreendida, por exemplo, na plotagem de um histograma.

Na plotagem de um histograma sem o agrupamento dos dados, visto que os dados são contínuos e podem não repetir na amostra, ou seja, resultaria em freqüência observada unitária a cada valor *x*. Imagine um histograma plotado utilizando freqüências iguais a um, seria uma reta paralela ao eixo *x*, não demonstrando o formato real dos dados.

De acordo com a apresentação em capítulos anteriores, o teste qui-quadrado de aderência trabalha com uma espécie de medida de distância conhecida por χ^2. No cálculo tradicional desta distância, somente os valores médios de cada classe são aproveitados.

Neste momento é válido notar uma desvantagem no método tradicional, no sentido de que toda informação da classe é representada por um único valor. Por exemplo, se o tamanho da amostra de dados (*n*) é igual a cem, tradicionalmente os dados seriam distribuídos em dez classes ($K = \sqrt{n}$ classes), sendo assim, somente dez valores médios de cada classe fariam parte do cálculo da distância χ^2, evidenciando a deficiência na representatividade dos dados, ou seja, de uma amostra de cem valores, apenas dez contribuem para tal cálculo. Na presente metodologia proposta, este problema não ocorre, pois a mesma trabalha sem o agrupamento dos dados em classes.

Devido à escolha de utilizar os dados não agrupados, três inovações no cálculo da distância χ^2 também se fizeram necessárias, as quais estão descritas a seguir:

A primeira inovação se dá em relação ao tamanho da amostra. Quando o tamanho da amostra de dados (*n*) cresce, a distância χ^2 também tende aumentar, constituindo evidências de rejeição do modelo teórico testado (tende para H_1). Este aumento observado na distância χ^2 ocorre devido ao grande número de parcelas em seu cálculo, mesmo na presença de pequenas dispersões. A sensibilidade de detecção de pequenas diferenças é uma característica estatística inerente a grandes amostras. Portanto, da amostra de dados observados (*n*) é extraída uma amostra de dados (*na*) para o cálculo da distância χ^2, buscando encontrar um equilíbrio entre a amostra de dados (*na*) e a distância χ^2, mantendo o máximo da informação dos dados.

Para exemplificar esta inovação, a Tabela 4.1 a seguir apresenta alguns tamanhos (*n*) de amostras e seus respectivos tamanhos (*na*) para o cálculo da distância χ^2. A matemática deste procedimento é demonstrada no próximo capítulo.

Tabela 4. 1 - Amostras para cálculo da distância χ^2, com grau de confiança de 95% e margem de erro de 2,5%.

n	na
10	10
20	20
30	29
40	39
50	48
60	58
70	67
80	76
90	85
100	94
150	137
200	177
300	251
400	317
500	377
600	432
700	481
800	526
900	568
1000	606
2000	869

Observe nesta tabela que para amostras (*n*) de tamanho pequeno, a metodologia considera praticamente quase todos os valores, porém, conforme as amostras (*n*) crescem, é selecionado parte desses dados (*na*) de forma sistemática.

A segunda inovação está relacionada aos graus de liberdade. Tradicionalmente, o teste qui-quadrado de aderência segue aproximadamente uma distribuição qui-quadrado com *K-1* graus de liberdade, onde *K* = número de classes. Devido a metodologia proposta nesta pesquisa, não trabalhar com classes, imagine que cada classe é representada por um único valor, a distribuição qui-quadrado então tem *n-1* graus de liberdade, mas conforme escrito anteriormente, da amostra de valores observados (*n*) é extraída uma amostra de dados (*na*), portanto, a distribuição segue *na-1* graus de liberdade.

Para concluir, a terceira inovação é referente ao cálculo da probabilidade de significância (valor-p). Verificou-se que conforme os graus de liberdade *na-1* aumentam, a distribuição qui-quadrado tende para uma distribuição normal, é por esse

motivo que nesta metodologia, a distribuição qui-quadrado é aproximada pela distribuição normal.

Sabendo que a função densidade da distribuição qui-quadrado tem uma maior exigência matemática, como cálculo de integrais, computacionalmente a distribuição normal torna-se mais vantajosa, pois sua função densidade é de mais fácil processamento. Porém, há uma desvantagem, a simetria da distribuição qui-quadrado está diretamente relacionada com o tamanho da amostra (n). Realizando alguns testes, verificou-se que a partir de trinta observações (Teorema Central do Limite) a distribuição qui-quadrado se aproxima da distribuição normal.

Capítulo 5

A Matemática da Metodologia Proposta, seus Resultados e Validações

Este capítulo apresenta o desenvolvimento matemático das inovações da metodologia proposta. Apresenta ainda os resultados de um experimento comparativo entre a metodologia proposta e o método tradicional do teste qui-quadrado de aderência.

5.1 A Matemática

O primeiro passo está relacionado à avaliação da dispersão dos dados em relação à média, ou seja, verificar a existência de pontos discrepantes. Para isto, utilize as medidas descritivas quartis e extremos, conforme descritas no capítulo 2. Em seguida, elabore a tabela de distribuição de freqüências, lembrando, sem o agrupamento dos dados. Calcule a distribuição de freqüências, incluindo a freqüência observada absoluta, acumulada absoluta e acumulada relativa.

O segundo passo é referente ao resumo dos dados (análise descritiva). Calcule o tamanho da amostra (n), a média aritmética amostral ($\bar{x}$), moda, mínimo, máximo, o desvio padrão (s) e a variância (s^2), conforme descrito pelo capítulo 2.

No terceiro passo, efetue os cálculos dos modelos teóricos de distribuição de probabilidades que se deseja testar. Utilizando a função de distribuição acumulada, calcule a distribuição de probabilidades de cada modelo. Para os modelos que não possuem a função de distribuição acumulada definida, por exemplo, normal e

lognormal, utilize o método de integração numérica, conhecido como Regra do Trapézio (LEITHOLD, 1994). A integração é realizada pelo fato de que a função de distribuição acumulada é obtida pela integral da função densidade de probabilidade do modelo.

No quarto passo, relacionado ao teste de aderência, calcule a distância χ^2 do teste qui-quadrado de aderência, conforme expressão 3.3 apresentada pelo capítulo 3. Mas antes de iniciar o cálculo da distância χ^2, efetue os seguintes cálculos de tamanho de amostra (COCHRAN, 1977) para o teste de aderência:

$$nc = \left(\frac{z}{me} \right)^2 * 0{,}25 \qquad (5.1)$$

onde:

nc = tamanho calculado,

z = coeficiente z relativo a um determinado grau de confiança,

me = margem de erro,

em seguida:

$$na = \frac{nc}{1 + \left(\frac{nc}{n} \right)} \qquad (5.2)$$

onde:

n = tamanho da amostra de dados,

na = tamanho da amostra para o teste da aderência,

nc = tamanho calculado,

e ainda:

$$va = \frac{n}{na} \tag{5.3}$$

onde:

va = valor de avanço entre as posições do vetor de dados,

na = tamanho da amostra para o teste da aderência,

n = tamanho da amostra de dados.

Iniciando o cálculo da distância χ^2, utilize como O_i o valor da freqüência acumulada absoluta e como E_i o valor da função de distribuição acumulada do modelo. Lembrando que o valor resultante da função de distribuição acumulada *F(x)* é sempre relativo. Sendo assim, é necessário transformá-lo para absoluto, da seguinte forma:

$$E_i = F(x) * n \tag{5.4}$$

onde:

n = tamanho da amostra de dados.

Primeiramente realize um somatório s_o_i para os primeiros valores O_i e um outro somatório s_e_i para seus respectivos E_i, faça isso enquanto a condição da expressão 5.5 a seguir for verdadeira:

$$s\_e_i < 5 \tag{5.5}$$

Para um posterior uso, quando s_e_i atingir o valor cinco, salve a posição de parada (*pp*) do vetor de um dos somatórios (s_o_i ou s_e_i). Assuma esses somatórios como a primeira parcela do cálculo da distância χ^2.

A segunda parcela do cálculo da distância χ^2 são os valores (O_i e E_i) dos vetores na posição:

$$pv = pp + va \tag{5.6}$$

onde:

pv = posição dos valores,

pp = posição de parada no vetor, quando s_e_i atingir o valor cinco,

va = valor de avanço entre as posições do vetor de dados.

As próximas parcelas são os valores de posição dada pela expressão 5.7 a seguir, sucessivamente até o final dos vetores.

$$pv = pv + va \tag{5.7}$$

onde:

pv = posição dos valores,

va = valor de avanço entre as posições do vetor de dados.

Observe neste quarto passo a seleção dos valores que fazem parte da distância χ^2, construindo a amostra (*na*) citada no capítulo 4.

O quinto passo é referente ao cálculo do valor-p. Efetue o cálculo do valor-p somente para os modelos que satisfaçam a condição da expressão 5.08 a seguir.

$$\chi^2 \leq l\_s \tag{5.08}$$

para:

$$gl = na - 1 \tag{5.09}$$

$$l\_s = \mu + (15 * \sqrt{\sigma^2}) \tag{5.10}$$

$$\mu = gl \tag{5.11}$$

$$\sigma^2 = 2 * gl \tag{5.12}$$

onde:

l_s = limite superior (encontrado heuristicamente),

gl = graus de liberdade,

na = tamanho da amostra para o teste da aderência,

μ = média,

σ^2 = variância.

Se a condição da expressão 5.08 não for satisfeita, rejeite o modelo teórico testado automaticamente, pois sua distância χ^2 ultrapassa o limite estipulado de quinze desvios padrão. Este limite foi encontrado após a realização de vários testes, iniciando por uma quantidade menor de desvios.

Se a condição da expressão 5.08 for satisfeita, divida a distância χ^2 em cem intervalos de delta *0,01*l_s*, assuma esses intervalos como os valores do eixo *x* e submeta-os à função densidade do modelo de distribuição normal, uma normal com os parâmetros média e variância apresentados pelas expressões 5.11 e 5.12 respectivamente.

Para calcular a área (*ar*), efetue um somatório da integração dos resultados gerados pela função densidade da normal (método de integração numérica, Regra do Trapézio). Por fim, obtenha o valor-p pela seguinte expressão:

$$valor\text{-}p = 1 - ar \qquad (5.13)$$

onde:

ar = área.

Para concluir, elabore as seguintes hipóteses:

H_0: não há diferença entre os dados observados e o modelo teórico testado,
H_1: há diferença.

Tendo como referência o valor-p, ordene os modelos testados, da melhor para a pior aderência aos dados observados. Lembrando que um valor-p pequeno, constitui evidência contra a hipótese H_0 (TRIOLA, 1999), ou seja, o modelo testado não representa os dados observados.

É importante destacar que nesta metodologia foi utilizado como critério de decisão o valor-p, mas poderia ser utilizado outro critério, por exemplo, o erro médio quadrático entre a distribuição dos dados e o modelo testado.

5.2 Metodologia Proposta versus Método Tradicional

É apresentada a seguir uma comparação entre a metodologia proposta na presente pesquisa versus o método tradicional qui-quadrado para teste de aderência.

Antes de iniciar a explanação do experimento comparativo, é importante destacar que esta metodologia foi implementada em um software, chamado Módulo de Aderência (MD), o qual está descrito no capítulo seguinte.

O experimento foi organizado da seguinte forma:

1) Para gerar as amostras com números aleatórios de cada modelo foi utilizado o software Input Analyzer 6.0 (INPUT, 2000);

2) Três tamanhos de amostras foram utilizadas: $n = 30$, $n = 200$ e $n = 2000$;

3) Para verificar a aderência foram utilizados os softwares MD, Input Analyzer 6.0 e Statistica 6.0 (STATISTICA, 2001);

4) O valor-p foi utilizado para mostrar qual dos três softwares melhor reconheceu o modelo de distribuição de probabilidade que os dados seguem. Ressaltando que estatisticamente utiliza-se a seguinte regra de decisão:

valor-p $> \alpha$ ➔ aceita H_0, isto é, aceita-se que os dados seguem o modelo testado;

valor-p $\leq \alpha$ ➔ rejeita H_0, isto é, afirma-se que os dados não seguem o modelo testado.

Os resultados desse experimento são apresentados na Tabela 5.1, considerando os distintos modelos de distribuição e um nível de significância $\alpha = 0,05$.

Tabela 5. 1 - Metodologia Proposta versus Método Tradicional

			valor-p			
Modelo (Parâmetros)	**Amostra (*n*)**	**Estimativas**	**MD**	**Input**	**Statistica**	**Melhor valor-p**
Uniforme (2 ; 12)	**30**	(2,13 ; 11,64)	0,0001	0,0545	0,0093	Input
		(2,88 ; 11,79)	0,2077	0,7500	0,0688	Input
		(2,49 ; 11,48)	0,9983	0,2560	0,1687	MD
	200	(2,04 ; 11,80)	1,0000	0,1500	0,8603	MD
		(2,03 ; 11,99)	0,0001	0,2570	0,0484	Input
		(2,03 ; 11,99)	1,0000	0,6000	0,5410	MD
	2000	(2,00 ; 11,99)	1,0000	0,7500	0,9064	MD
		(2,00 ; 11,98)	1,0000	0,7500	0,6208	MD
		(2,01 ; 11,99)	1,0000	0,6380	0,3872	MD
Exponencial (6)	**30**	(6,83)	0,9969	0,5700	0,5167	MD
		(5,71)	0,9849	0,0050	0,9288	MD
		(6,17)	0,9996	0,3780	0,8700	MD
	200	(6,56)	1,0000	0,6790	0,2103	MD
		(6,30)	1,0000	0,4860	0,6390	MD
		(5,83)	1,0000	0,3180	0,1824	MD
	2000	(6,17)	1,0000	0,7110	0,9109	MD
		(6,23)	1,0000	0,7420	0,2306	MD
		(5,95)	1,0000	0,4460	0,7406	MD
Normal (8 ; 2)	**30**	(7,95 ; 1,87)	0,9761	0,0050	0,0217	MD
		(7,35 ; 1,68)	0,9991	0,0050	0,2093	MD
		(8,37 ; 1,71)	0,9991	0,0050	0,7120	MD
	200	(7,97 ; 1,87)	1,0000	0,2670	0,5034	MD
		(7,88 ; 1,84)	1,0000	0,6850	0,3400	MD
		(8,15 ; 1,97)	1,0000	0,6820	0,1768	MD
	2000	(8,04 ; 1,95)	1,0000	0,1940	0,1914	MD
		(7,97 ; 1,99)	1,0000	0,3670	0,2672	MD
		(7,97 ; 2,06)	1,0000	0,6740	0,9208	MD
Lognormal (0,6 ; 1)	**30**	(-1,01 ; 1,14)	0,9991	0,0050	0,0616	MD
		(-1,18 ; 0,82)	0,9635	0,0050	0,7173	MD
		(-1,05 ; 1,12)	0,0404	0,0050	0,0001	MD
	200	(-1,04 ; 1,18)	1,0000	0,0050	0,5849	MD
		(-1,27 ; 1,18)	0,9999	0,5790	0,8214	MD
		(-1,14 ; 1,10)	0,0001	0,0194	0,1253	Statistica
	2000	(-1,21 ; 1,12)	1,0000	0,6380	0,4164	MD
		(-1,12 ; 1,16)	1,0000	0,1100	0,4456	MD
		(-1,20 ; 1,16)	1,0000	0,0476	0,0506	MD
Triangular (4 ; 10 ; 16)	**30**	(5,18 ; 9,99 ; 13,81)	0,0849	0,0651	não faz	MD
		(4,70 ; 10,42 ; 14,76)	0,6188	0,2810		MD
		(5,92 ; 8,98 ; 12,55)	0,0001	0,2870		Input
	200	(4,63 ; 9,28 ; 15,54)	0,0001	0,0230		Input
		(4,44 ; 10,55 ; 15,48)	0,6774	0,7500		Input
		(4,79 ; 9,30 ; 15,20)	0,0001	0,5040		Input
	2000	(4,31 ; 9,95 ; 15,83)	0,9998	0,2480		MD
		(4,06 ; 10,16 ; 15,70)	0,0001	0,3780		Input
		(4,19 ; 9,81 ; 15,99)	1,0000	0,2270		MD

Analisando a tabela, observa-se que em todos os casos dos modelos exponencial e normal, o software MD reconhece o padrão dos dados. Em relação aos modelos uniforme, lognormal e triangular, o MD falha em alguns casos.

As células destacadas de cinza, demonstram os casos em que houve divergências entre os softwares testados. Após uma análise criteriosa desses casos divergentes, observa-se que as flutuações aleatórias no processo de geração dos dados influenciam nas estimativas dos parâmetros. O software MD por trabalhar com os dados individuais (sem agrupamento), mostra-se alta sensibilidade em relação às estimativas dos parâmetros dos modelos.

Esta alta sensibilidade é notória nos casos do modelo triangular, o qual possui três parâmetros e o mesmo apresentou uma quantidade maior de rejeição do modelo, ou seja, valor-p $\leq 0,05$. Em geral, os modelos possuem um ou dois parâmetros. Por outro lado, nos casos em que o MD não reconhece o padrão triangular, ele sugere o modelo normal, o que é aceitável.

É possível verificar que na grande maioria dos modelos testados, o software MD apresenta o melhor valor-p. Vale lembrar que um valor-p muito pequeno constitui evidências de rejeição do modelo testado.

Além das comparações em relação à aderência dos modelos, esta seção também destaca a validação do procedimento do software MD para detecção de pontos discrepantes. Comparou-se as medidas quartis e extremos apresentadas pelo MD em relação às apresentadas pelo software Statistica.

Capítulo 6

Implementação do Software

Conforme previamente descrito pelo capítulo anterior, um software chamado Módulo de Aderência (MD), foi implementado com base na metodologia proposta por esta pesquisa.

Este capítulo descreve a especificação formal do software MD e suas interfaces. O software MD foi desenvolvido na linguagem de programação visual C++ (BORLAND, 2002).

6.1 Especificação Formal do Software

Nesta seção o software MD é descrito formalmente pela Linguagem de Modelagem Unificada (UML). A UML é uma notação, principalmente diagramática, para modelagem de sistemas, usando conceitos orientados a objetos (LARMAN, 2000).

O primeiro diagrama a ser apresentado é o Diagrama de Caso de Uso. Um diagrama de caso de uso é um documento narrativo que descreve a seqüência de eventos de um ator (um agente externo) que usa um sistema para completar um processo.

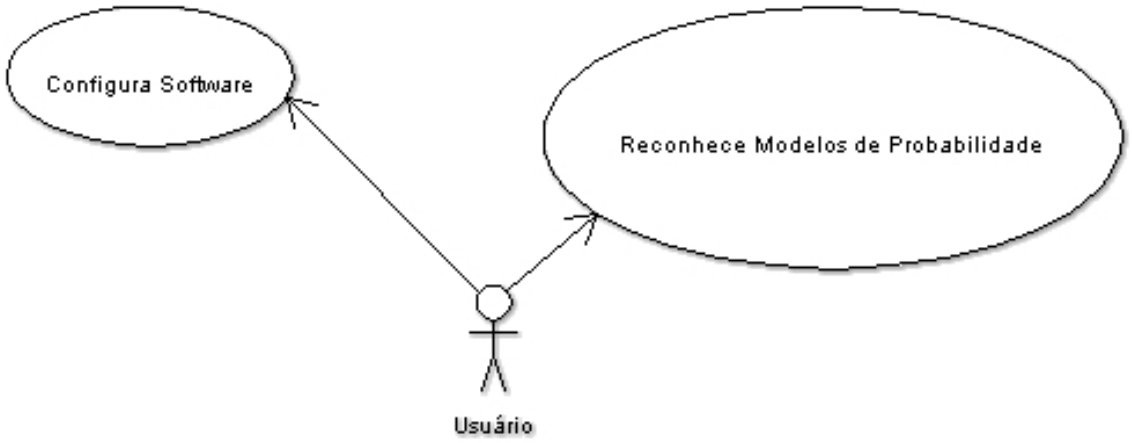


Figura 6. 1 - Diagrama de casos de uso.

Observe que no diagrama mostrado pela Figura 6.1, existem dois casos de uso, Configura Software e Reconhece Modelos de Probabilidade. Tendo em vista que a metodologia desta pesquisa é aplicada no caso de uso Reconhece Modelos de Probabilidade e para um melhor entendimento deste caso de uso, o mesmo será detalhado na forma de um Diagrama de Estado apresentado pela Figura 6.2 a seguir.

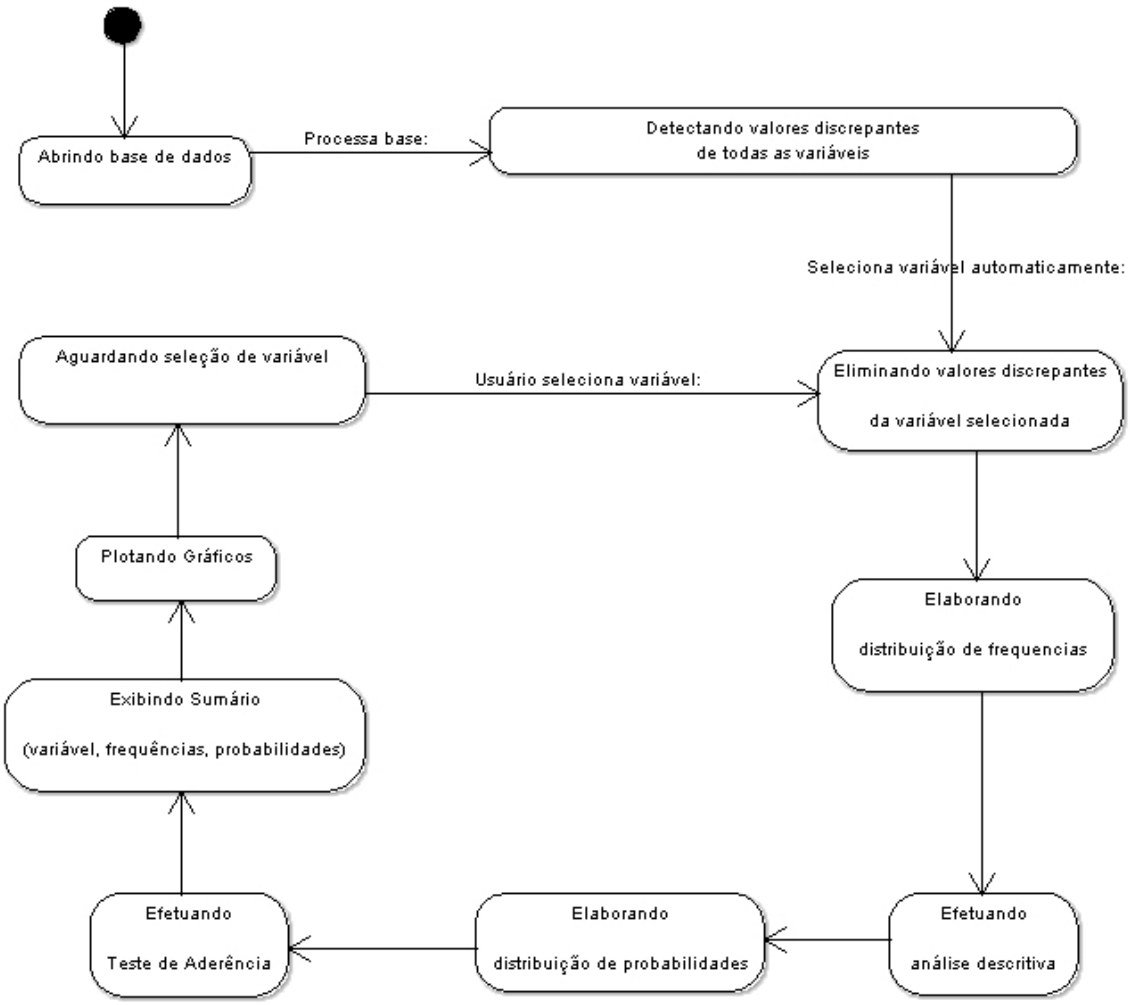


Figura 6. 2 - Diagrama de estado do caso de uso Reconhece Modelos de Probabilidade.

Os diagramas de estado mostram o ciclo de vida de um objeto, os eventos pelos quais ele passa, as suas transições e os estados em que ele está entre estes eventos (LARMAN, 2000).

Um terceiro e último diagrama a ser apresentado é o Diagrama de Classes, o qual ilustra as especificações para as classes de software e de interfaces de uma aplicação. As informações típicas são: classes, associações, atributos, tipos de dados, interfaces, métodos e dependências.

A Figura 6.3 a seguir apresenta o diagrama de classes do software MD.

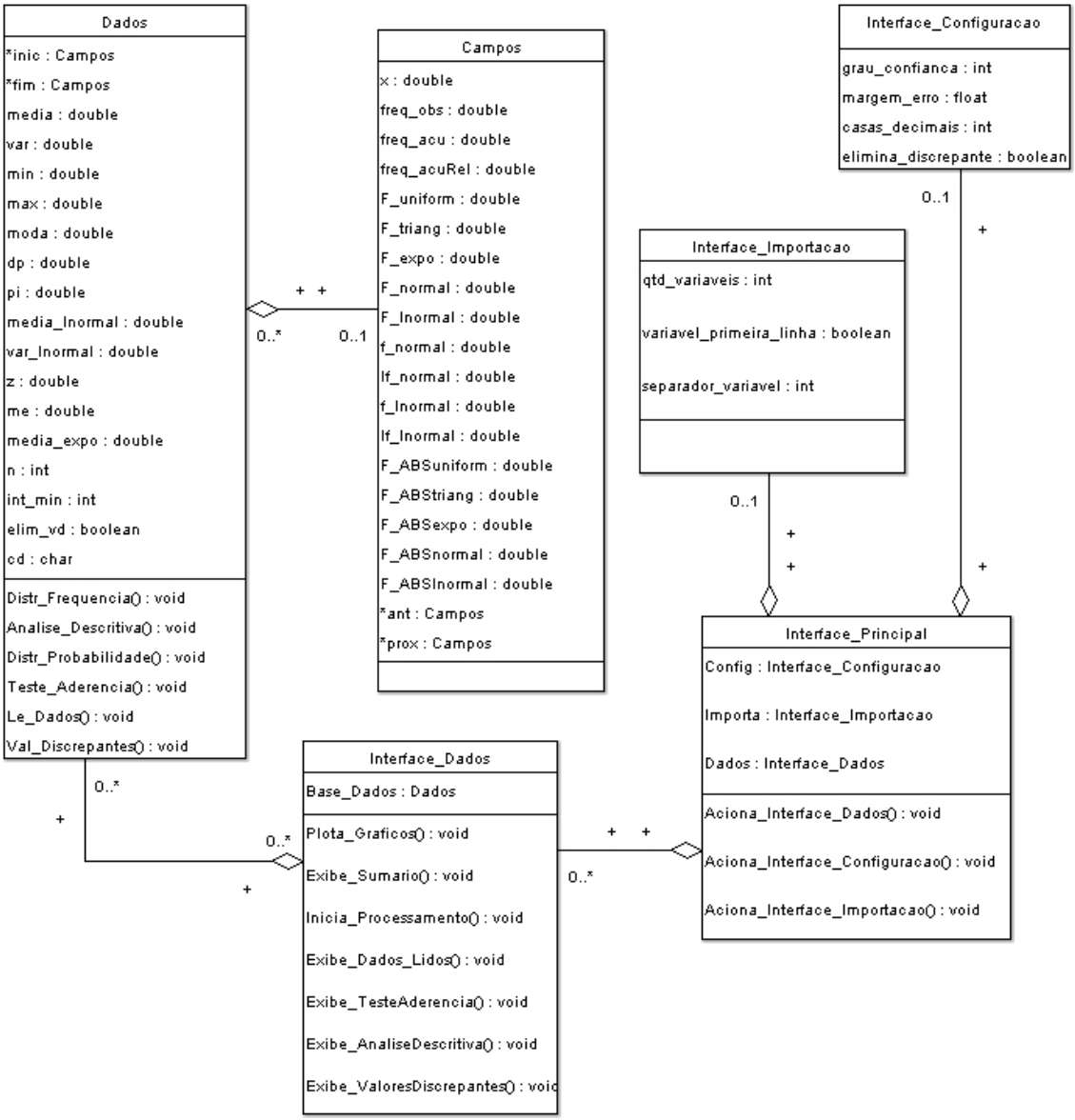


Figura 6. 3 - Diagrama de classes do software MD.

6.2 Interfaces do Software

Pela interface de Configuração apresentada na Figura 6.4 a seguir, o usuário pode alterar configurações do software como Grau de confiança, Margem de erro, Casas decimais e se o usuário deseja que o software elimine ou não os pontos discrepantes.

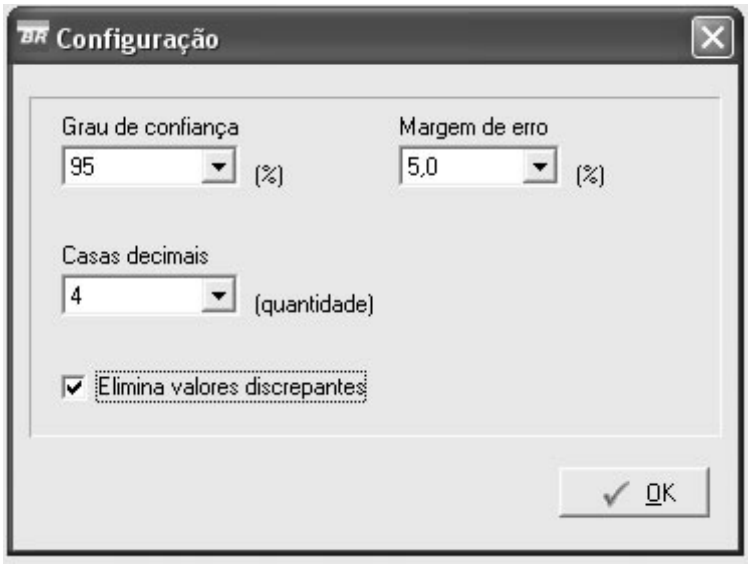


Figura 6. 4 - Interface de Configuração.

O software MD faz o acesso aos dados que obrigatoriamente precisam estar dispostos em arquivo de formato texto, similarmente a uma tabela, onde nas colunas estão representadas as variáveis e nas linhas os dados observados.

Caso esse arquivo contenha mais de uma variável, essas variáveis precisam estar separadas por algum separador aceito pelo algoritmo de importação dos dados implementado no software. Os nomes das variáveis podem estar na primeira linha do arquivo, caso não estejam, o software atribui os rótulos $V_1, V_2, ..., V_k$ para cada variável, onde k = número de variáveis.

A Figura 6.5 apresenta a interface do software que interage com o usuário no sentido de obter a quantidade de variáveis no arquivo, se os nomes das variáveis estão na primeira linha e caso o arquivo contenha mais de uma variável, qual o separador das variáveis.

Figura 6. 5 - Interface de importação dos dados.

A interface da Figura 6.5 é exibida ao usuário solicitando algumas informações, esta interface é exibida logo após a escolha do arquivo a ser analisado. O usuário informa as solicitações da interface e clica no botão OK, os dados são importados e todo processo de reconhecimento de modelos probabilísticos é automaticamente iniciado.

Em seguida é apresentada ao usuário uma interface dividida em algumas guias, são elas, Dados, Teste de aderência, Gráficos, Sumário e Valores discrepantes. A Figura 6.6 apresenta a guia Dados, esta guia exibe as variáveis e os dados importados do arquivo.

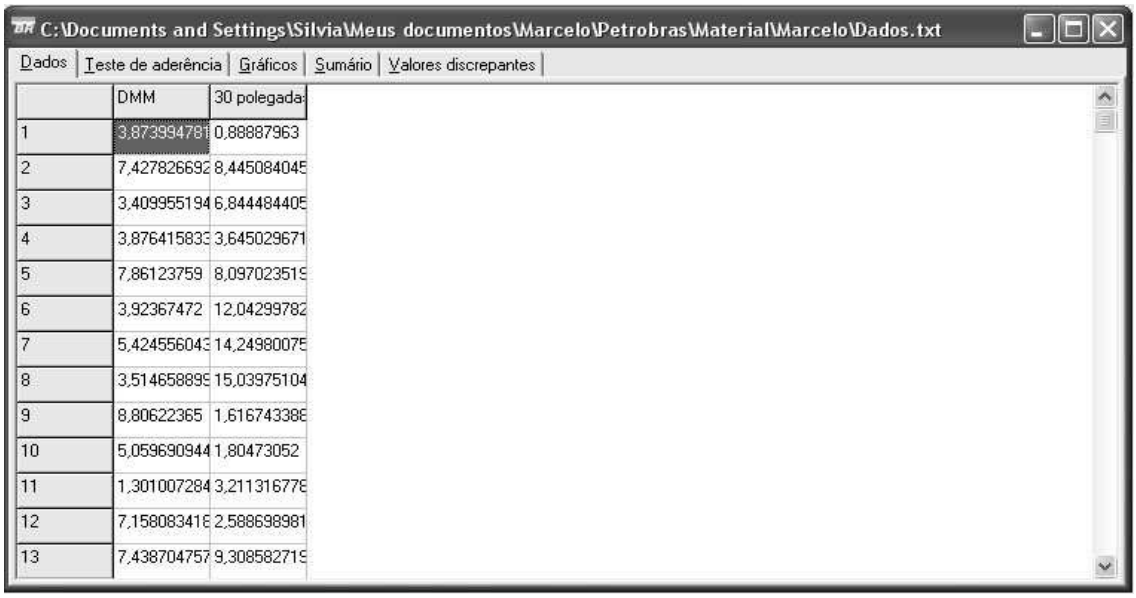


Figura 6. 6 - Guia Dados.

Na guia Dados o usuário somente visualiza as variáveis e os dados, não é permitida nenhuma alteração de conteúdo.

Todas as guias ficam disponíveis de modo que o usuário possa navegar por elas na ordem que lhe convém. A Figura 6.7 apresenta a guia Teste de Aderência.

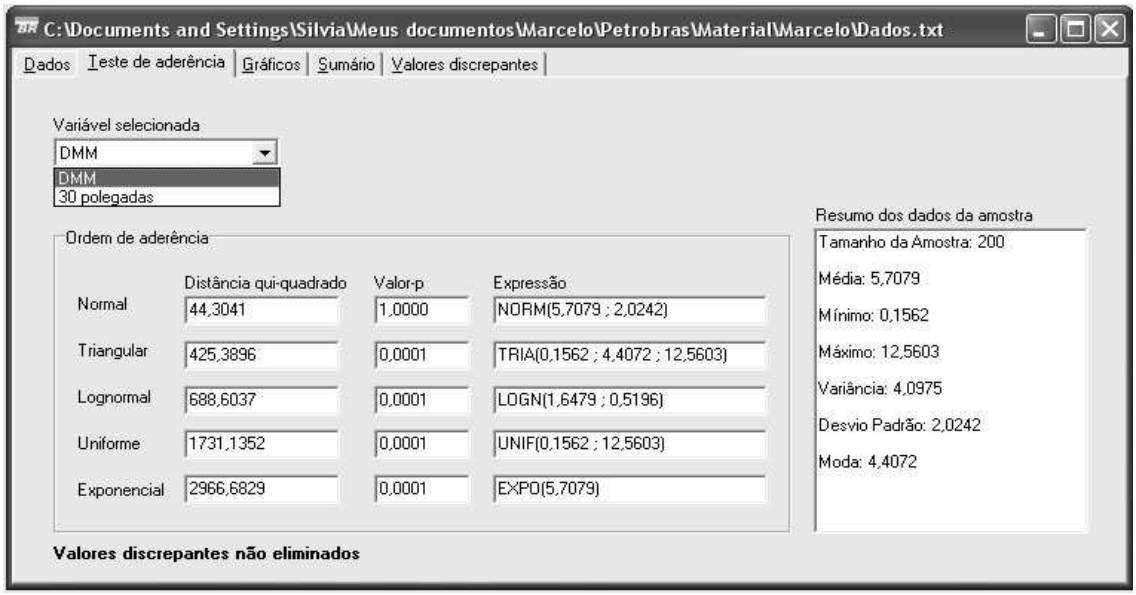


Figura 6. 7 - Guia Teste de aderência.

Conforme pode ser observado pela Figura 6.7, a guia Teste de Aderência disponibiliza ao usuário os seguintes itens:

1) Seleção da variável a ser analisada;
2) Resumo da amostra, contendo tamanho da amostra, média, mínimo, máximo, variância, desvio padrão e moda;
3) Seqüência da melhor para a pior aderência dos dados aos modelos testados, informando a distância qui-quadrado, valor-p e as expressões de cada modelo. Os modelos testados são: exponencial, triangular, normal, lognormal e uniforme;
4) Exibe uma mensagem informando se os valores discrepantes foram ou não eliminados.

No primeiro item, seleção da variável, o usuário pode escolher uma outra variável a ser analisada, após esta escolha, o software efetua novamente todo processo de reconhecimento de modelos para esta nova variável selecionada.

No terceiro item, seqüência da aderência, tendo como referência o valor-p, o software apresenta os modelos em ordem crescente de qualidade na aderência.

A próxima guia chamada de Gráficos, apresenta um gráfico para cada modelo testado, como pode ser observado pela Figura 6.8 a seguir.

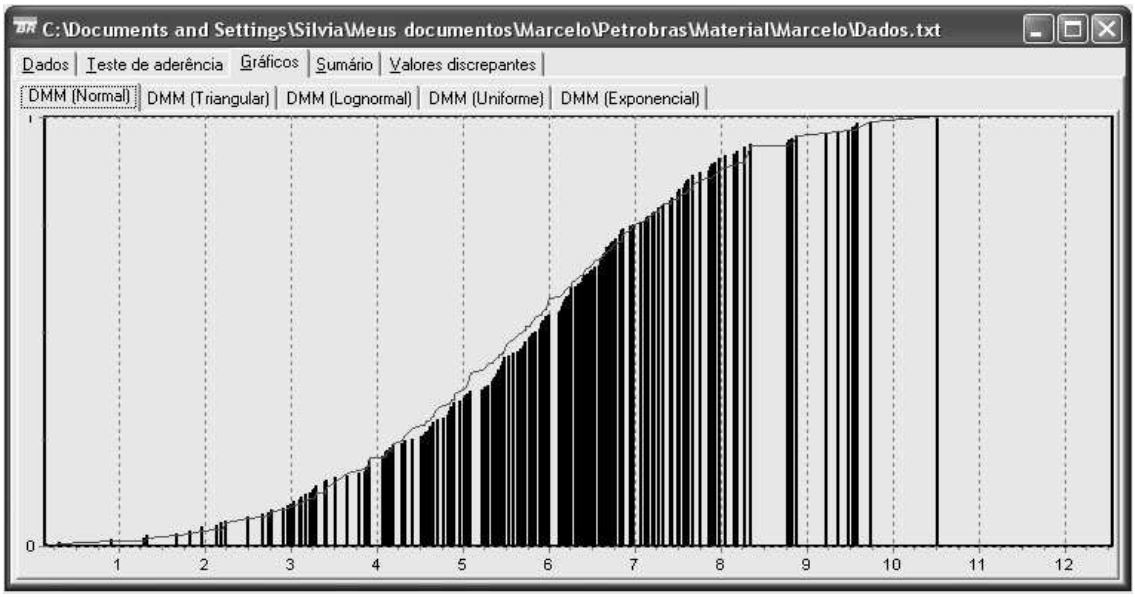
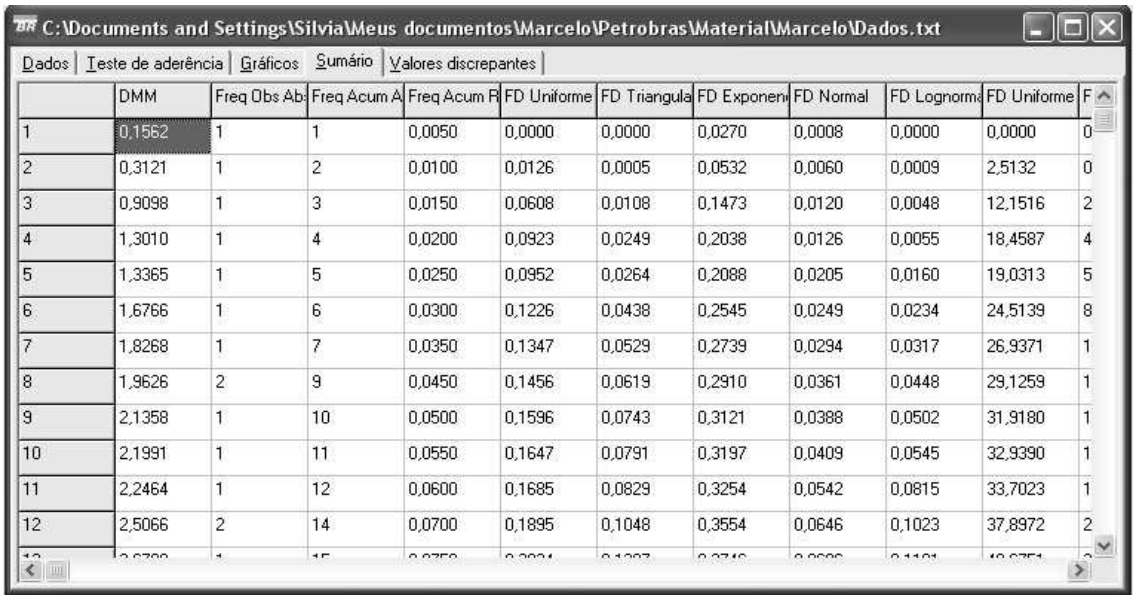


Figura 6. 8 - Guia Gráficos.

As barras verticais representam o histograma da distribuição de freqüências relativas. A curva que acompanha o histograma, representa a função de distribuição acumulada do modelo teórico. Observe que existem várias guias de nome composto pelo nome da variável analisada e entre parênteses o nome do modelo teórico testado, assim o usuário pode verificar graficamente se houve aderência ou não.

A Figura 6.9 apresenta a guia chamada Sumário, esta guia exibe os resultados dos cálculos efetuados para cada valor observado, tais como, distribuições de freqüências e de probabilidades acumuladas.


C:\Documents and Settings\Silvia\Meus documentos\Marcelo\Petrobras\Material\Marcelo\Dados.txt

Dados | Teste de aderência | Gráficos | Sumário | Valores discrepantes

	DMM	Freq Obs Abs	Freq Acum A	Freq Acum R	FD Uniforme	FD Triangula	FD Exponen	FD Normal	FD Lognorma	FD Uniforme	F
1	0,1562	1	1	0,0050	0,0000	0,0000	0,0270	0,0008	0,0000	0,0000	0
2	0,3121	1	2	0,0100	0,0126	0,0005	0,0532	0,0060	0,0009	2,5132	0
3	0,9098	1	3	0,0150	0,0608	0,0108	0,1473	0,0120	0,0048	12,1516	2
4	1,3010	1	4	0,0200	0,0923	0,0249	0,2038	0,0126	0,0055	18,4587	4
5	1,3365	1	5	0,0250	0,0952	0,0264	0,2088	0,0205	0,0160	19,0313	5
6	1,6766	1	6	0,0300	0,1226	0,0438	0,2545	0,0249	0,0234	24,5139	8
7	1,8268	1	7	0,0350	0,1347	0,0529	0,2739	0,0294	0,0317	26,9371	1
8	1,9626	2	9	0,0450	0,1456	0,0619	0,2910	0,0361	0,0448	29,1259	1
9	2,1358	1	10	0,0500	0,1596	0,0743	0,3121	0,0388	0,0502	31,9180	1
10	2,1991	1	11	0,0550	0,1647	0,0791	0,3197	0,0409	0,0545	32,9390	1
11	2,2464	1	12	0,0600	0,1685	0,0829	0,3254	0,0542	0,0815	33,7023	1
12	2,5066	2	14	0,0700	0,1895	0,1048	0,3554	0,0646	0,1023	37,8972	2

Figura 6. 9 - Guia Sumário.

Para finalizar, a Figura 6.10 apresenta a guia Valores discrepantes, a qual apresenta os valores discrepantes detectados em cada variável contida no arquivo.

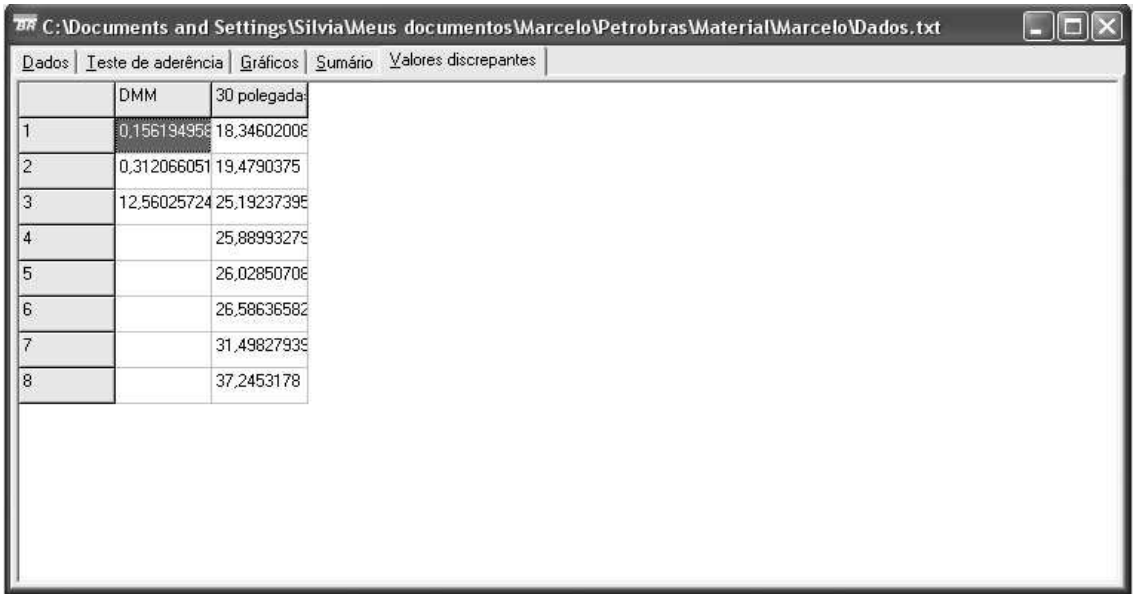


Figura 6. 10 - Guia Valores discrepantes.

Capítulo 7

Considerações Finais

Este capítulo apresenta as conclusões desta pesquisa, a qual tem como objetivo propor uma metodologia, alternativa às tradicionais, para o reconhecimento de modelos de distribuição de probabilidade. O capítulo destaca também algumas sugestões para futuros trabalhos.

7.1 Conclusões

A presente pesquisa está sustentada por uma base conceitual para exploração de dados, tendo como principal objetivo o reconhecimento de modelos probabilísticos. Entre os conceitos descritos, encontram-se técnicas de organização dos dados, como por exemplo, elaboração de tabelas de freqüências, medidas descritivas, distribuições de probabilidades e por fim a pesquisa descreve sobre testes de aderência.

A pesquisa detalha o estudo de uma metodologia, alternativa ao método tradicional, para reconhecimento de modelos probabilísticos utilizando o teste qui-quadrado de aderência.

A metodologia proposta possui as seguintes características:

1. Trabalha sem o agrupamento dos dados. Neste sentido, a metodologia não sofre algumas desvantagens que o método tradicional apresenta, por exemplo, deficiência na representatividade dos dados (valores médios de cada classe) e a

dificuldade de especificar uma quantidade satisfatória de classes a ser utilizada no agrupamento dos dados.

2. Utiliza as distribuições na sua forma acumulada. Esta característica tornou-se necessária pelo fato da forma acumulada permitir trabalhar com os valores individuais.

3. Realiza a crítica automática dos dados para identificação de valores discrepantes. Este é um diferencial importante, visto que os parâmetros dos modelos são estimados por dados observados, e a estimação é prejudicada com presença de valores discrepantes nos dados.

4. Calcula a probabilidade de significância (valor-p) por aproximação da distribuição qui-quadrado pela normal. Neste sentido, a metodologia tem uma vantagem matemática computacional, pois a função normal é de mais fácil processamento do que a qui-quadrado, por outro lado, há uma desvantagem em relação à simetria, a mesma pode estar deficiente em amostras pequenas.

5. Possui uma seleção sistemática de valores para o cálculo da distância χ^2. Além da vantagem da metodologia não sofrer dificuldades no agrupamento dos dados, a seleção sistemática dos dados que participarão da distância χ^2 também é um grande diferencial. Esta seleção busca encontrar um equilíbrio entre a amostra para aderência (*na*) e a distância χ^2, mantendo o máximo da informação dos dados.

O software Módulo de Aderência (MD) foi implementado com base na metodologia proposta. Testes comparativos foram realizados com o objetivo de validar a sensibilidade do MD em relação a outros softwares que também realizam o reconhecimento de modelos usando o método tradicional de aderência pelo teste qui-quadrado.

Observou-se um excelente desempenho que o MD teve em relação aos outros softwares para as distribuições normal, exponencial, lognormal e uniforme. Porém, o

MD apresentou limitações em relação à distribuição triangular, supõe-se que isto se deve ao fato das estimativas dos parâmetros desta distribuição.

7.2 Trabalhos Futuros

Após um amplo estudo que a presente pesquisa apresenta, é válido que ela também sugira algumas continuações, para que assim, o privilégio de criar novas idéias de alguma forma se perpetue.

A seguir são apresentadas algumas sugestões para trabalhos futuros:

A primeira sugestão de trabalho é o desenvolvimento de uma metodologia qualitativa para o reconhecimento de modelos de distribuição de probabilidade, a partir da elicitação do conhecimento de um especialista do domínio dos dados. A idéia é que o software ao interagir com o especialista, através de uma espécie de conversa, apresente o padrão desejado pelo especialista. O principal norte desta abordagem é eliminar a necessidade de dados observados para reconhecer os modelos probabilísticos.

A segunda sugestão está relacionada à metodologia desta pesquisa. Sugere-se validar a presente metodologia para outras distribuições contínuas de probabilidade, estendê-la para as distribuições discretas e para distribuições multivariadas.

Sugere-se como um terceiro trabalho futuro, estudos relacionados à estimação de parâmetros. Tendo em vista a grande sensibilidade que a presente metodologia demonstrou na estimação dos parâmetros dos modelos a serem testados.

Referências

ANDERSON, T. W.; SCLOVE, S. L. **An Introduction to the Statistical Analysis of Data.** 2nd edition. Palo Alto: Scientific Press, 1986.

AURÉLIO Eletrônico Século XXI, versão 3.0: Lexikon Informática Ltda, 1999.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. **Estatística Para Cursos de Engenharia e Informática.** São Paulo: Editora Atlas S.A., 2004.

BARBETTA, P. A. **Estatística Aplicada às Ciências Sociais.** 5ª edição. Florianópolis: Editora da UFSC, 2003.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms.** New York: Plenum Press, 1987.

BITTENCOURT, G. **Inteligência Artificial: ferramentas e teorias.** 2ª edição. Florianópolis: Editora da UFSC, 2001.

BORLAND C++ Builder, version 6.0: Borland Software Corporation, 2002.

BOWKER, A. H.; LIEBERMAN, G. J. **Engineering Statistics.** 2nd edition. New Jersey: Prentice-Hall, 1972.

CAZORLA, I. M. **Inferência Estatística.** Disponível em: http://www.socio-estatistica.com.br/Edestatistica/V. Inferencia Estatistica.doc. Acesso em: 20 de outubro de 2004.

COHEN, P. R. **Empirical Methods for Artificial Intelligence.** Cambridge: MIT Press, 1995.

COCHRAN, W.G. **Sampling techniques.** 3nd edition. NewYork: John Wiley, 1977.

COMPUTAÇÃO BRASIL. Sociedade Brasileira de Computação, Porto Alegre, ano V, edição 16, 2004.

INPUT Analyzer, version 6.0: Rockwell Software Inc, 2000.

JAIN, R. **The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling.** New York: John Wiley & Sons Inc, 1991.

JANKAUSKAS, L.; MCLAFFERTY, S. **Bestfit, Distribution Fitting Software by Palisade Corporation.** In: Winter Simulation Conference, Newfield, 1995.

LARMAN, C. **Utilizando UML e padrões: uma introdução à análise e ao projeto orientados a objetos.** Porto Alegre: Bookman, 2000.

LAW, A. M. **Simulation Modeling and Analysis.** 2nd edition. New York: McGraw-Hill, 1991.

LEITHOLD, L. **O Cálculo com Geometria Analítica.** 3ª edição. São Paulo: Harbra Ltda, 1994.

MEYER, P. L. **Probabilidade: Aplicações à Estatística.** 2ª edição. Rio de Janeiro: LTC, 1984.

MOORE, D. S.; MCCABE, G. P. **Introdução à Prática da Estatística.** 3ª edição. Rio de Janeiro: LTC, 2002.

MONTGOMERY, D. C.; RUNGER, C. G. **Estatística Aplicada e Probabilidade para Engenheiros.** 2ª edição. Rio de Janeiro: LTC, 2003.

NEWENDORP, P. D.; SCHUYLER, J. R. **Decision Analysis for Petroleum Exploration.** 2 nd edition. Aurora: Planning Press, 2000.

NIST/SEMATECH. **e-Handbook of Statistical Methods.** Disponível em: http://www.itl.nist.gov/div898/handbook/index.htm. Acesso em: 17 de janeiro de 2005.

PQRS, version 3.0: Sytse Knypstra, 1998.

ROMEU, J. L. **The Chi-Square: a Large-Sample Goodness of Fit Test.** START: Selected Topics in Assurance Related Technologies, Rome, v.10, n.4, p.1-6, 2003.

ROMEU, J. L. **Anderson-Darling: A Goodness of Fit Test for Small Samples Assumptions.** START: Selected Topics in Assurance Related Technologies, Rome, v.10, n.5, p.1-6, 2003.

ROMEU, J. L. **Kolmogorov-Smirnov: A Goodness of Fit Test for Small Samples.** START: Selected Topics in Assurance Related Technologies, Rome, v.10, n.6, p.1-6, 2003.

STATISTICA, version 6.0: StatSoft Inc, 2001.

SISA – Binomial, version 2.0: Daan Uitenbroek, 1997. Disponível em: http://home.clara.net/sisa. Acesso em: 18 de janeiro de 2005.

TRIOLA, M. F. **Introdução à Estatística.** 7ª edição. Rio de Janeiro: LTC, 1999.

U-CARE, version 2.02: Group Biometry and population Biology, 2003. Disponível em: ftp://ftp.cefe.cnrs-mop.fr/biom/Soft-CR/U-CARE/. Acesso em: 18 de janeiro de 2005.